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

Felianka Kaftandjieva

EALTA
www.ealta.eu.org

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

Felianka Kaftandjieva

EALTA | European Association for Language Testing and Assessment

ISBN: 978-90-5834-104-4

Cito, Arnhem, 2010

Foreword

When news about Dr. Felianka Kaftandjieva's untimely passing away reached EALTA's members in September 2009, there were many expressions of shock and of a profound sense of loss on the members' list. It became very evident, if any evidence was needed, that she was highly valued for her unstinting and generous support of EALTA in making presentations, in running workshops and in chairing EALTA's Membership Committee. Several members had also enjoyed her invaluable services as consultant to their development work on tests and examinations.

Of the several proposals the Executive Committee received that could honour her memory and her contributions to EALTA, it was felt that Felly herself would have appreciated any effort to spread knowledge and expertise to EALTA's members and to the assessment and testing audience in general. She had a passionate interest in teaching and believed that knowledge should be passed on openly, actively and generously.

With this as the starting point, and as a first step along other possible actions to honour Felly's memory, the Executive Committee decided to translate into English Felly's recent "big doctorate" that she defended successfully at the University of Sofia in 2008. She considered this as a summary of much of the work she had been doing on standard setting over more than a decade. During that period she applied several existing standard setting methods but also developed and tested several news methods. The PhD dissertation is a critical examination of these new methods.

EALTA is indebted to several members who have assisted in this endeavour. Plamen Mirazchiyski, a former graduate student of Felly's, translated the text into English as a pro bono homage to his mentor. He did the job quickly and on top of his demanding work at the IEA Data Processing and Research Center in Hamburg. As editors of the volume and during the first check of psychometric terms, we were able to assess Plamen's thorough work, for which we are very grateful. We are also grateful to Jamie Dunlea and Charles Alderson who did a further revision of the text from both the linguistic and content point of view with a very tight deadline.

EALTA is also indebted to Cito for agreeing to publish this second EALTA publication.

We wish to acknowledge, in particular, the assistance of José Noijons in managing the publication process.

We hope that this publication marks the beginning of a set of activities that makes Felly's work better known among our profession. We believe that her work richly deserves this recognition.

Sauli Takala
Norman Verhelst

Editors

Table of content

	Introduction	**7**
1	**Content standards, standard setting, criteria for success and cut scores**	**9**
1.1	Research question	9
1.2	Basic terms in the theory and methodology of standard setting	12
1.3	Major stages in the development of standard setting methodology	17
2	**Methods for setting cut scores**	**29**
2.1	Review of methods	29
2.1.1	General description of methods	29
2.1.2	Classification schemes	30
2.1.3	Main characteristics of methods for setting cut scores	34
2.2	An illustrative example	39
2.2.1	Instrument	40
2.2.1.1	Test description	41
2.2.1.2	Psychometric characteristics of the test	41
2.2.1.3	Performance standards	44
2.2.2	Judgment	47
2.2.2.1	Assigning the judgment task	47
2.2.2.2	Judges	49
2.2.2.3	Intra-judge consistency	50
2.3	Setting cut scores	61
2.3.1	Basket procedure	61
2.3.2	Compound Cumulative method	67
2.3.3	Cumulative Cluster method	71
2.3.4	ROC-curve method	74
2.3.5	Item Mastery method	80
2.3.6	Level Characteristic Curve method	85
3	**Comparative analysis of the quality of the separate methods**	**91**
3.1	Methods	91
3.1.1	Goal, main purposes, object, subject and hypotheses of the current research	91
3.1.2	Study design	92
3.1.2.1	Instruments	93
3.1.2.2	Judgments	95
3.1.2.3	Resampling	97
3.2	Empirical results	98
3.2.1	Cut scores	98
3.2.2	Replicability and precision of the cut scores	102
3.2.2.1	Degree of matching in resampling	102
3.2.2.2	Standard error of the cut scores	103

3.2.3	Commensurability of the cut scores	107
3.2.3.1	Common tendencies	107
3.2.3.2	Significance of the differences between cut scores X and Y	109
3.2.3.3	Significance of the differences between the different methods	110
3.2.4	Classification consistency	111
3.2.4.1	Replicability of the classification decisions	111
3.2.4.2	Equivalence of the classification decisions	114
3.3	Comparative analysis of the methods	122
3.3.1	Criteria for quality	122
3.3.2	Evaluation of the quality of the methods	126
	Conclusions and recommendations	**131**
	References	**137**
	Appendices	**155**
1	List of the existing methods for setting cut scores	156
2	Illustrative example: test items and judgment	159
3	Illustrative example: Table for transformation of the raw test score into Z-scale and cut scores	160
4	Common European Framework of Reference for Languages: Descriptors for reading	161
5	ROC-curve method: Indices of sensitivity and specificity and optimization criteria	163
6a	Cut scores for the different judges and methods (Raw test score)	164
6b	Cut scores for the different judges and methods (Z-scale)	165
6c	Cut scores – descriptive statistics	166
7	Degree of consistency between the classification decisions	167
8	Criteria for quality: Pair-wise comparison of the analysed methods for setting cut scores	168

Introduction

In the last two years the interest in achievement testing appears to be increasing (once again) in Bulgaria. Unfortunately, as in any beginning, enthusiasm prevails over professionalism. Support for such a negative statement is the fact that one of the most complex and discussed problems in achievement testing – that of setting cut scores in criterion-referenced tests – is completely ignored, as if it were not a relevant topic.
The significance of this problem is determined by the direct dependence of test score interpretation on the cut scores being set. If the cut scores are inadequate they raise serious doubts about the validity of the interpretation of the test results.
The complexity of the problem is created by a series of other problems – theoretical, methodological and practical – but probably the major one is the great variety of methods (currently over 60) for setting cut scores. This variety of methods complicates the choice of the most appropriate method for particular test situations. An additional complication is the fact that many of the existing methods lack sufficient evidence and arguments to support their validity.
This problem applies to several methods which have been used extensively in foreign-language testing in Europe in the last few years. That is why the **main research question** that this study is aiming to answer is: *Which of the selected six methods for setting cut scores that have been developed for criterion-referenced tests is the most efficient* with regard to (a) practical applicability and (b) internal validity?
The six methods, the **object of the current study**, are: the Basket procedure, the Compound Cumulative method, the Cumulative Cluster method, the ROC-curve method (Receiver Operating Characteristic), the Item Mastery method and the Level Characteristic Curve method.
A detailed description of the **methods used in the current study** can be found in Chapter Two (2.1), and the goal, object, subject and the three research hypotheses are presented in Chapter One where the formulation of the research question is presented (1.1).
In **Chapter One**, besides the formulation of the research question, a short review of the basic terms and stages in the development of the standard setting methodology is presented.
In **Chapter Two**, after a review of the existing methods, a detailed description of the six methods is presented. The description of each is based on illustrative examples in which the emphasis is on the specifics of the judgment typical of this kind of method. The methods of determining the degree of consistency between the judgments and the empirical data are also reviewed and a detailed description is provided of the new index of consistency (MPI, misplacement index) developed by the author especially for the needs of this type of judgment.In addition, the problem of the misplacement of the cut scores that can be expected in this kind of judgment is also analyzed.
Chapter Three is devoted to the comparative analysis of the quality of the different methods. This comparative analysis is based on the results of three empirical studies in which the six methods for defining the cut scores were used. The main focus of the comparative analysis is concerned with: (a) the precision (in terms of standard error) of the

obtained cut scores; (b) the significance of the differences between two consecutive cut scores and (c) the degree of consistency of the classification decisions made on the basis of the different methods. The results from the comparative analysis provide sufficient evidence supporting each of the three research **hypotheses**. They confirm that:

- The method with minimal standard error of the cut score is the Compound Cumulative method;
- The Basket procedure differs substantially from the other five methods for setting cut scores;
- The Compound Cumulative and the Cumulative Cluster method result in almost equivalent cut scores.

The choice of the **most effective** method for setting cut scores is based on the application of six criteria for evaluating the quality of the methods that were developed especially for this purpose. In developing these criteria, a balance between internal and procedural validity was sought. Account was taken of both the criteria proposed in the research literature as well as the context of the concrete situation. The final decisions about the most effective method for setting cut scores is made through pair comparison and analysis of the obtained individual profiles. As a result of this analysis, the **Compound Cumulative method** emerges as the most effective method. This finding provides the answer to the **main research question** and the goal of the study is achieved.

The seven **conclusions** and the eight **recommendations** deriving from them are logical consequences and generalization of the study results.

The study itself could not have been accomplished without the active participation and support of:

- the institutions which organized and conducted the tests and judgments;
- the authors of the particular tests that were used in the research;
- the participants who took the tests whose results provided the empirical data;
- the judges that took part in the process of making judgments.

I owe a special debt of gratitude to the Department of Social Education at the Faculty of Primary and Pre-school Education – Sofia University "St. Kliment Ohridski" for the support and the opportunity for one-year sabbatical, which resulted in the current monograph.

The list of people who helped me one way or another over the years is too long for me to acknowledge each one of them personally. Nevertheless, in this list the names of two of my colleagues, friends and co-authors must be mentioned: Professor Sauli Takala (Finland) and Professor Norman Verhelst (The Netherlands), to whom I would like to express personally and publicly my deep gratitude for the long and heated arguments, the critical comments, the intellectual challenge, the full and unreserved support and the beneficial collaboration in the last ten years.

1 Content standards, standard setting, criteria for success and cut scores

1.1 Research question

One of the major goals of the Bulgarian National Program for Development of the School Education and Pre-school Education and Training (2006-2015) is *"building up an effective system for internal assessment through widespread use of tests and introducing a system for national standardized external assessment"* (National Program, p. 15).

The main arguments for *"overall introduction of tests"* as a major instrument for internal and external assessment according to the National Program (2006, p. 15) is that they:

- will ensure **objective measurement** of the educational outcomes;
- will make possible the **comparability** of the results between students from different schools and groups of graduates from different school years.

Concerning the planned annual external assessment at the end of each stage of education (grades 4, 7, 10, 12) through national standardized exams, according to the National Program (2006, p. 16) there are two major goals:

1 Ascertaining the level of coverage of the State Educational Requirements (Content Standards) for the corresponding stage;
2 Using the results from the graduating examination in "basic education" (comprising primary and lower secondary grades 1-7) and general secondary education (at grade 12) for entering the next stage of education.

The goals of the external assessment at the end of each stage of education, formulated in this way, make it clear that the tests for the national standardised exams will be **criterion-referenced** and the role of criteria will be played by the State Educational Requirements (which are actually Content Standards) for the particular stage of education. At the same time at least two of the exams (at the end of grades 7 and 12) will have a **norm-oriented** element because the students will be compared with each other on the basis of their exam results and these results will be used for selective purposes in applying for various types of schools.

Such a combination of the two approaches (criterion-oriented and norm-oriented) for marking tests is possible in principle (Nitko, 1980; Linn, 1994; Nitko, 1994) and is sometimes applied in achievement testing. This approach is, in fact, used for the assessment and interpretation in one of the few standardized achievement tests – the test for reading comprehension of PIRLS (see Stoyanova, 2004, pp. 252-254).

Irrespective of the approach chosen (criterion-referenced, norm-referenced or combined), the obligatory stages in constructing any measuring instrument (including tests) are:

- Setting standards for the interpretation of the test results, as well as
- Setting the relevant cut scores in the particular measurement scale used in considering the test results.

Different from the measurement itself, which in the case of multiple-choice items can be maximally "objectified", setting the standard and the corresponding cut scores is not and

cannot be objective because the evaluation itself is subjective by nature.
The outcome of measurement, whatever it is, is by itself neither good (high) nor bad (low). It becomes so only in a comparison with some standard determined in advance. If this standard is defined "by eyeballing" (approximately), by a directive from the Ministry of Education or by a scientific method, nothing changes the fact that standard setting does not exist objectively (external to and independent from our consciousness). 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.
Indeed, the subjectivity of standard setting as well as the importance of the decisions to be taken on the basis of these standards, which affect the future of the students tested, are among the main reasons that make the setting of cut scores one of the most complex, contradictory and controversial problems in the area of achievement testing (Berk, 1986, p. 137; Zieky, 1994, p. 29; Hambleton, 1998, p. 103). The only means for minimizing the dispute according to Cizek is *"crafting 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, p. 46-47).
The fact that Cizek calls for the development of methods in setting cut scores does not mean that there is a lack of such methods. More than twenty years ago, in guiding the development of cut scores Berk discussed thirty-eight such methods (Berk, 1986).
Today, there are over sixty methods (see Appendix 1), and there is a tendency for new methods to emerge at regular intervals.

There are several reasons for this methodological variety:

- Firstly, no one method is universal, i.e. applicable in each test situation. This means that in each particular test situation it is necessary to carefully choose the most appropriate method and – if necessary – to modify it or to create another method.
- Secondly, the theory of achievement testing is developing and this leads to the creation of new item formats and new forms of evaluating and summarizing test results. This, in turn, poses the necessity of modifying the existing methods and of creating new methods for setting cut scores.
- Thirdly, the lack of sufficiently deep knowledge of the existing methods often leads to "reinventing the wheel", i.e. creation of "new" methods, which are to a large extent replications of the already existing ones (Reckase, 2000b, p. 2).

Irrespective of the reasons for this variety, in each particular test situation, one of the major problems is choosing the method for setting the cut scores. This problem is complicated by the fact that different methods applied to the same test situation lead to different cut scores and to different classification decisions concerning the achievement of the test takers (Jaeger, 1989; Bontempo et al, 1998 and others). Taking into account how important the consequences of such a choice are, one of the main criteria for method selection is the availablity of sufficient eidence of its validity. This criterion is claimed to be the major one not only by leading experts in the field (Hambleton & Pitoniak, 2006; Norcini & Shea, 1997; Cizek, 1996), but it is also included, in one way or another, in all professional codes and

quality standards in the area of testing. For instance, The International Test Commission recommends (2.7.9): "Use passing scores (cut scores) in test interpretation only when evidence of the validity for the pass scores is available and supports its use." (htpp://www.intestcom.org)

Unfortunately there are numerous cases of methods being used for setting cut scores whose validity is not confirmed by adequate scientific research and publications.

A typical example of this is the method for setting cut scores recommended by the Council of Europe in its pilot version of a Manual which provided guidelines on how foreign language examinations can be related to the Common European Framework of Reference for Languages[1] (Council of Europe, 2003, pp. 89-97). This method was applied for the first time in the second phase (year 2001) of the international project for internet-based foreign language testing DIALANG (http://www.dialang.org/intro.htm) and subsequently in other national projects in Europe (for example see Noijons & Kuijper, 2006).

At the same time as this method was being recommended, new methods for setting cut scores were being developed (Kaftandjieva, et al, 1999; Verhelst & Kaftandjieva, 1999; Kaftandjieva & Verhelst, 2000; Kaftandjieva & Takala, 2002), which went through field tests and were applied in several international and national European projects for linking the results from foreign language tests to the standards that were presented in the CEFR. These methods, although applied successfully, are not sufficiently familiar to the broad testing community and there is insufficient documentation of the evidence of their validity.

All these methods are intended for the setting of cut scores (single or multiple) in criterion-referenced tests for reproductive skills, including mainly multiple-choice items or open-ended items requiring short answers. In the field of foreign language testing such item types are typical of tests of listening and reading comprehension as well as tests of grammatical and vocabulary knowledge. Such is also the intended format for most of the national standardized exams in Bulgaria at the end of each educational stage, at least as can be seen from the performance standards (known in Bulgarian as "education-examination programs") for state maturity exams (Ministry of Education and Science, 2006).

Bearing in mind the key role that the methods for setting cut scores play in interpreting test scores and in decision making, it is obvious that empirical data and their analysis are necessary to support their validity and advisability. In any case, comparative data and criteria for the evaluation of the different methods' quality are necessary to facilitate the choice of a particular method in a particular situation.

In other words, the **main research question** that the current study will try to answer is: *Which one of the six selected methods developed by the author for setting cut scores that have been developed for criterion-referenced tests is the most efficient with regard to (a) practical applicability and (b) internal validity?*

The main goal of the study is *to describe six methods developed by the author for setting cut scores in criterion-referenced tests and provide relevant arguments (theoretical and empirical) for supporting their validity and, on the basis of comparative empirical analysis, to determine the most effective among them for mass use.*

1 Henceforth referred to with the acronym CEFR

Objects of the current study will be the following methods for setting cut scores:

1 *The Basket procedure*
2 *The Compound Cumulative method*
3 *The Cumulative Cluster method*
4 *The ROC-curve method*
5 *The Item Mastery method*
6 *The Level Characteristic Curve method*

The subject of the current study will be the psychometric characteristics of the aforementioned methods related to their validity.

During the development and application of the aforementioned methods several **hypotheses** were delineated and will be tested in the study. These hypotheses are as follows:

I The method having minimal standard error of the cut score is the **Compound Cumulative method**.
II The cut scores derived using the **Basket procedure** differ substantially from those obtained by the remaining methods in the current study, and the direction of the difference will depend on the position of the particular cut score on the scale used for the measurement of the test results.
III The methods producing cut scores with values that are closest to each other will be the **Compound Cumulative** and the **Cumulative Cluster method**.

In testing these hypotheses as well as in the description of the six methods and their main characteristics it is necessary to use a specific methodology and terminology. Hence the presentation of the results from the empirical research will be preceded by an introduction to the main terminology and the theoretical foundations of the methodology for setting cut scores.

1.2 Basic terms in the theory and methodology of standard setting

Standards, Cut scores and Linking

In the field of achievement testing, standard setting is regarded as a process of decision making concerning the classification of the test/exam results in several, successive, but limited number of levels of achievement *(achievement, proficiency, mastery, competency)*. The levels of achievement themselves are called **performance standards**[2] and are clear and precise definitions of what exactly a person taking the test has to do to demonstrate competency in the particular content area (CRESST, 1999). As it can be seen from the definition, the performance standards are inextricably linked with some concrete content area and the respective **content standards**. In other words, while the content standards state WHAT a given test taker has to know and be able to do, the performance standards define **TO WHAT EXTENT** the test taker has to know and be able to do what the corresponding content standards state.

From everything stated so far it is clear that for setting the performance standards it is

2 Editors' note: This is used synonymously with the term Performance Level Descriptions, PLDs

not enough to fix the number of levels (e.g. five) and give them the corresponding labels (eg., A – excellent, B – above average, C – average, D – usually the minimum passing grade, and F – fail).

The other obvious conclusion is that even if the number of levels in different content areas is equal, their meaning and interpretation would be different depending on the concrete content domains themselves. For instance the CEFR (2001 pp. 35-37) differentiates six basic levels for language proficiency which, regardless of the particular skill (reading, writing, listening, etc.), have the same names and labels (A1 – Breakthrough, A2 – Waystage, B1 – Threshold, B2 – Vantage, C1 – Effective Operational Proficiency, C2 – Mastery), but depending on the concrete content they mean different things. For example, if person X has level B2 in watching TV and movies this would mean that he or she can understand most of the news and broadcasts such as documentaries, interviews, discussions, plays and most movies if the language and pronunciation are standard (CEFR, 2001, p. 93).

The same level – B2 – concerning correspondence means that if a given person is at that level, then he or she *"Can write letters conveying degrees of emotion and highlighting the personal significance of events and experiences and commenting on the correspondent's news and views"* (CEFR, p. 106). Thus, it is expected, and practice confirms, that the same person can be on a different level (i.e. covering different performance standards) in the different areas of competency even when it is in the same content domain – e.g. command of a particular foreign language.

The direct link between the performance standards and particular content area made Hansche (1998, p. 4) define the performance standards as a **system** that includes:

- *performance levels* – numbers and names;
- *performance descriptors* – concrete descriptions of what exactly each person on the respective level must know, be able to do and demonstrate;
- *benchmarks* – illustrative examples of items and responses that are characteristic of each performance level;
- *cut scores* – for each assessment scale in which the test scores are expressed and which distinguish two consecutive performance levels.

This four-component system of performance standards in practice operationalises also the respective content standards and allows us to specify and update them.

Closely related to this systematic approach is Kane's (1994, p. 426) concept of the cut score. According to this concept, the cut score is the operationalised version of the respective level of competence, while the performance standard is the conceptual version of the same level. In other words the **cut scores** are those points of the scale used for representing the test results, each one distinguishing two consecutive levels of competence (performance standards).

From everything written until now it follows that:

Firstly, the cut scores have no independent meaning and are directly bound to (a) each concrete test/exam and the corresponding assessment scale and (b) the corresponding performance standards.

Secondly, the performance standards themselves are dependent on the corresponding content domain and on the concrete content standards for this domain and stage of education.

To designate this **linking** between the cut scores (and correspondingly the tests) and the performance standards, which in turn are linked to the content standards, the English term *"alignment"* is frequently used. According to the CRESST (1999) glossary, *alignment* refers to the process of linking content and performance standards to instruction, and assessment and learning in classrooms. Narrowing the range of the term *alignment*, Linn (2001) defines it as *"... the degree to which assessments accurately reflect standards"* (Linn, 2001, p. 5).

The analysis of these definitions leads to the logical conclusion that, on the one hand, linking the cut scores, performance standards and standard setting has a direct connection with all aspects of the validity of the assessment. On the other hand, the setting of the cut scores, as an inherent part of this linking, plays a central role in determining the meaning of, and subsequently the interpretation of, the test results and as such is at the core of each argument for the validity of the assessment (Dylan, 1996, p. 288).

Put another way, an inadequate or incompetent setting of cut scores will lead to invalid test results regardless of the reliability or validity of the test itself.

Setting standards and/or cut scores is directly connected with another term that is used widely in the English-language literature – namely the term of *linking*. Differently from alignment which is concerned mainly with the content aspects of the instruction and assessment, linking is mainly concerned with the correspondence between the different scales used to represent the test results obtained by two different tests or other kinds of measurement instruments. Depending on the characteristics of the two instruments being compared, different models of linking can be used. According to the current classification (Mislevy, 1992; Linn, 1993), there are five major categories that the different models of linking can be associated with: *equating, calibration, projection, statistical moderation and social moderation*.

The methods for setting cut scores which are the object of this research belong to the last category of this classification scheme – **social moderation**. As the name of the category suggests, in this kind of linking the subjective assessment in the form of judgments is taken into account to establish the correspondence between the two measurement instruments. In the case of setting cut scores, correspondence between the results of a given test and the respective performance standards will be sought. The performance standards themselves represent an ordinal assessment scale. Despite the necessity of (subjective) judgments, it is typical of social moderation that in this kind of linking there are no defined limits on the characteristics of the two tests. Concerning the models from the other categories of the classification scheme, there are numerous preliminary and quite rigorous requirements concerning both tests/instruments that are the object of linking.

In the context of the Bulgarian educational practice, it is obvious that creating tests and conducting annual national standardized exams at the end of each educational stage is only the first step in the long series of steps on the road towards achieving the main goals of this overall introduction of the tests into the educational system: ensuring the comparability of the test results of students from different schools and groups of graduates, determining the level of coverage of the state standard settings for the corresponding stage and using the test results for entrance into the next stages.

Concerning foreign language education (in Bulgaria and elsewhere in Europe) the situation

is slightly more optimistic because the performance standards are already set in the CEFR. Moreover, thanks to several international projects, a cluster of illustrative examples for the separate levels is already available and these illustrative samples are constantly being updates and added to.

Standard setting strategies

Nevertheless, there is another problem in foreign language testing that is still not acknowledged in Bulgaria. This is the problem of choosing a strategy for generalizing results of several tests/exams/assessments. This problem appears in the case of more than one test/exam when for each one of them a system of performance standards and the respective cut scores is already created. As a result a test taker is assigned a score for each test or assignment, and very often these results are represented in different measurement scales with a different number of categories. Sometimes these results have to be generalized in one way or another and expressed in a single generalized mark.

The procedures used for combining the results from several different measurements are called *standard setting strategies*.

Although not the only possible ones, there are three most frequently applied standard setting strategies for generalizing the results from different measurements in education: the *compensatory*, *conjunctive* and *mixed* strategies.

The **compensatory strategy**, as its name suggests, allows a high level of competence in one of the components of the assessment to compensate for a low level of the other components. The general score in this strategy is either the sum of the results from the different components or their mean. The compensatory strategy is based on the assumption that the sum of the separate components reflects adequately the measured construct. Nevertheless, in any single case this assumption needs theoretical grounding and empirical evidence. As the summarised result usually has a higher reliability than the separate components, in the absence of other considerations this strategy is to be recommended rather than the other two strategies (Haladyna & Hess, 1999; Hambleton et al., 2000; Hansche, 1998).

The **conjunctive strategy** requires attaining some predefined minimum level of competence for each one of the separate components to allow the final, summarized result to be judged as acceptable (sufficient). Regardless of the fact that the reliability of this strategy is likely to be lower than the one obtained in the compensatory strategy, there are concrete situations in which its application is relevant. For instance in order to obtain a driving licence each applicant has to pass both parts of the exam (theoretical and practical) while a high score in either part does not compensate a potentially low score in the other. The reason for applying this strategy in this concrete example is that each driver has not only to drive the car, but also has to know the respective laws and rules. Of course, applying the conjunctive strategy suggests that sufficiently high reliability is guaranteed for each component of the exam.

The **mixed strategy** is applied in cases in which some of the components of the general score are assumed to be more important than others. In this case, for such components, some minimum level of competence is required, while for some of the rremaining components, high levels are allowed to compensate for low levels in other components.

Each of the aforementioned strategies has its advantages and disadvantages. The choice would depend on the concrete situation and the comparative analysis of the possible application of the separate strategies.

Absolute and relative norms

Setting cut scores is historically related with the appearance and development of **criterion-referenced testing**, although cut scores (norms) exist in norm-referenced testing, and in practice it is possible to create a criterion-referenced test without setting cut scores, i.e. points on the assessment scale by which the test results are to be represented in a limited (smaller) number of assessment categories.

The major difference between cut scores in norm- and criterion-referenced tests is in the methods used for their setting and their great variety. Especially in criterion-referenced tests, this variety is considerable compared to the methods in setting cut scores (norms) in norm-referenced tests.

The standards used in criterion-referenced tests are often defined as **absolute** while in norm-referenced tests they are considered to be **relative**. This classification of performance standards is based on the articles of the following authors:

- Nedelsky (1954), who is the author of the first method for setting cut scores that bears its author's name. The article is entitled "Absolute Grading Standards for Objective Tests";
- Glaser (1963), who first introduced the term *criterion-referenced* measures and defined this type of measurement as measurement depending on *"absolute standards for quality"*, different from the norm-referenced measures which are based on relative standards (p. 519).

Unfortunately this classification, and the related terminology, are not quite precise because:

Firstly, the so-called absolute standards are also relative in their nature because the achievement of a given test taker is being compared, not with the results of other test takers, but with the range of the respective content domain and with what the test taker had to learn and demonstrate during the measurement.

Secondly, absolute standards, as well as absolute truth, do not exist. In fact, the performance standards in criterion-referenced testing are based mainly on judgments. This judgment is made by persons whose opinion is to a great extent dependent on their experience. In turn, this experience suggests that the experts have accumulated the normative information which, although not openly, is influencing their judgment. That is why, instead of denying the presence of a normative component in defining the *"absolute"* performance standards used in criterion-referenced testing in recent years (Zieky, 1994, p. 31; Norcini, 1994, p. 169; Kane, 1994, p. 453; Hambleton, 1998, p. 104; Downing et al., 2006, p. 51), there is a clear tendency in:

a purposeful delivery of normative information for those who are included in the judgment task;
b using of the normative information for the validation of the judgments or – put in another way – for a *reality check*.

Concerning the performance standards used for decision making in the interpretation of

the results from norm-referenced tests, they are also subject to judgment. For example the minimal standard score for membership in MENSA (an international association of people with a high intelligence quotient, founded in 1946) is a test result over the 98th percentile, and not 97 or 99, which is clearly subjective and is not based on any objective grounds. Other international organisations of people with a high intelligence quotient had, for one reason or another, adopted other minimal performance standards (e.g. for Intertel this is the 99th percentile; for the International High IQ Society it is the 95th percentile; for both Prometheus and the Triple Nine Society it is the 99.9th percentile).

1.3 Major stages in the development of the standard setting methodology

As was noted above, setting a standard and fixing the relevant cut scores is a process of decision-making. With or without purposefully applying specific methods, individuals make decisions and evaluate themselves, other people or things around them every day by classifying them as good or bad, high or low quality, important or unimportant and make their choices which in turn can be evaluated as adequate or not. That is why a number of authors search for the roots of the methodology for setting standards far back in antiquity, giving examples from ancient China and Egypt and the Old Testament (Green, 2000; Zieky, 2001).

Zieky (1994) distinguishes **four** major stages in the development of standard setting and cut score methodology which he defines as the ages of innocence, awakening, disillusionment and the age of realistic acceptance.

During the long **age of innocence** that ends in the 1960's, almost no attention was paid to how the standards and the cut scores were set. This, of course, does not mean that they did not exist, but that they were not an object of systematic scientific research and development. According to Zieky (1994, p. 4), the most frequently used methods were those which are based **on tradition** *(The cut score will be 70% correct item responses because it has always been like this)* or **on authority** *(The cut score will be 60% correct item responses because I think it is)*. One more method can be added to these, which can be described as **the Goldilocks method** *(The cut score will be 80% correct item responses because 70% is too little and 90% is too much)*.

The next age – the **age of awakening** – is mainly related to the emergence and development of criterion-referenced testing and is accompanied by a sizeable number of publications related to the creation of new methods for setting cut scores as well as an in-depth analysis of the basic characteristics and the corresponding quality of the different methods.

This can be seen in the graph in Figure 1, which shows the number of scientific publications in the educational literature (in English only) related to criterion-oriented testing (left side of the graph) and setting cut scores (right side of the graph) in the period between 1971 and 2006. The left side of the graph is a replication of Hambleton's research. In his article "The Rise and Fall of Criterion-Referenced Measurement?" Hambleton (1994) analyses the scientific publications in the English literature devoted to criterion-referenced testing, published in the period between 1971 and 1991. Both studies use the same bibliographic

database - ERIC (Education Resources Information Center: http://www.eric.ed.gov/), but while Hambleton analyzes a more narrow period in time focusing only on publications related to criterion-referenced testing, the current study broadens the scope and the period and adds publications related to standard setting and cut scores.

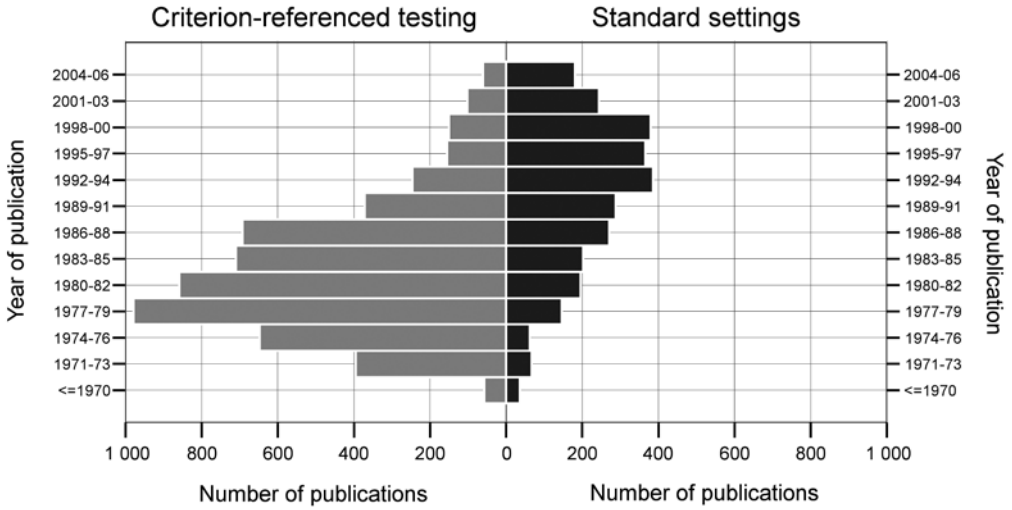


Figure 1 Scientific publications for the period 1971–2006

As can be seen from the graph, the number of scientific publications related to criterion-referenced testing reaches its maximum at the end of 1970's and decreases after that. Nevertheless, interest in the methodology of setting standards and cut scores is still evident as most of these publications are located in the 1990's.

This persistent interest in the problems of standard setting and cut scores can be explained, in part, with the complexity and debates in the domain. On the other hand, it can be related to the world-wide tendency for increasingly stronger relations between the three major components of the educational process: content standards, process of instruction, and assessment. This, in turn, leads to the appearance and development of new forms of measurement and assessment *(alternative assessment, authentic assessment, performance assessment, curriculum-based assessment, standard-based assessment, and portfolios)* and, respectively, shifting the interest from criterion-referenced testing towards these new forms. Nevertheless, these new forms, although having different names, are in essence criterion-referenced because they are aimed towards the measurement and evaluation of what students know and are able to do (Hambleton, 1994, p. 22). In these new forms of measurement and evaluation performance standards are needed, if the existing methods are not adjusted to cope with the new test formats. This leads to the creation of new methods or the modification of current ones, which has kept the interest in these problems alive for more than thirty years.

A crucial moment in the development of standard setting and cut scores is the publication

of Glass's article (1978), which directs serious criticism towards the entire process of setting cut scores, defining them as *"blatantly arbitrary"* (Glass, 1978, p. 258), and appeals for maximal limitation of the arbitrary manner and randomness in decision-making. Subsequently this appeal has been repeated by numerous leading specialists in this area (Shepard, 1980; Zieky, 2001; Linn, 2003 and others.).

Although Glass was almost stigmatized by the scientific community because of his strong criticism, his article had a considerable effect on the subsequent development of standard setting (cut score) methodology and especially on the problems of validation and the quality criteria for different methods.

Glass's article sets the beginning of **the age of disillusionment**. According to Zieky (1994), the two periods – of awakening and disillusionment – overlap as *"Some authors appear to be living in the Age of Awakening and others in the Age of disillusionment."* (Zieky, 1994, p. 16).

A key moment in the age of disillusionment is the realization that no one method can change the fact that standard setting is subjective in nature. In fact, the subjective nature of standard setting is "the apple of discord" in the area of achievement testing and this is rather strange because in the theory of decision-making this has been known for a long time without provoking such wild discussions. There are three possible explanations for the reasons that gave rise to the discussions about the essence of standard setting in education:

Firstly, it seems that striving for absolute truth is deeply rooted in every human being. According to Goldman (1999) *"the desire for truth occupies a central role in workaday cognitive practices such as magic, divination and religion"* in all cultures irrespective of their technological development (Goldman, 1999, p. 32).

Secondly, setting cut scores, which is usually done after the judgment has taken place, includes complex computational procedures whose only purpose is to generalize and quantify the opinions of the individual experts who took part in the judgment task. In this way, however, the subjective nature of the evaluation is being masked *"which in turn gave the entire process a patina of professionalism and propriety"* (Cizek, 2001, p. 7). In other words, awe of numbers and the fact that the cut scores were fixed with the aid of a computer (i.e. "objectively") and not by human beings (i.e. "subjectively"), have an important role in the interpretation of these cut scores.

Thirdly, decision-making in everyday life usually affects a limited number of people while standard setting and the respective cut scores are of great importance, because they not only influence the future of the test takers, but also serve as a basis for numerous other administrative decisions related to the management of the educational system. In other words, setting standards is an administrative decision, and as such, can become an object of criticism by those who are not satisfied withthe decisions or the processes. That is why according to Cizek (2001),*"Standard setting is perhaps the branch of psychometrics that blends more artistic, political, and cultural ingredients into the mix of its products than any other"* (Cizek, 2001, p. 5).

The transition from the ages of awakening and disillusionment towards the **age of realistic acceptance** starts in 1983 when, according to Zieky, setting cut scores had been transformed from *"an esoteric topic limited to psychometricians or statisticians"* into a topic

covered in every textbook in educational testing (Zieky, 2001, p. 26).
Conformity to existing realities leads to some changes in the methodology for setting standards, which are conditioned mainly by changes in the methodology of testing, namely:

- Changes in the test item format and the more extensive use of items with extended constructed responses, which involve a transition from dichotomous (correct/incorrect) scale to polytomous scales for scoring the responses to items.
- Changes in the evaluation and interpretation of the test results, which also moves from dichotomous (e.g. pass/fail) into more fine-grained gradation, i.e. becomes polytomous (e.g. beginner – intermediate – advanced).
- Changes related to a more extensive application of Item Response Theory (IRT) in test construction and the analysis of test results, which is increasingly replacing classical test theory.
- Changes in the practice of test administration and the increasing use of computer-based tests, including adaptive and on-line testing.

All these changes in the methodology of testing impose changes in the methodology of setting cut scores for the resulting tests. An additional intrinsic reason is related to the accumulated experience in the domain of setting cut scores and its rationalisation.
In the mid 1980's, it became obvious that different methods for setting cut scores, even when applied to one and the same test, led to different values. Summarizing the results of twelve comparative studies, Jaeger (1989) analysed 32 pairs of cut scores (in terms of the number of correctly answered test items) obtained by applying two different methods for one and the same test. For each one of these 32 comparisons, he calculated the ratio of the higher cut score to the lower and found that this ratio varies in too broad a range (between 1 and 52). Moreover, in almost half of the cases (15 out of 32), one of the cut scores is almost 1.5 times higher than the other (Jaeger, 1989, p. 500).
Whether the difference between two cut scores is big or small becomes apparent only when it is compared with the standard error of measurement. For instance, the standard error of the measurement of a good test with average difficulty and consisting of 50 items is usually about three points on the measurement scale (Harvill, 1991, p. 35). Let us imagine that for setting cut scores for this test we applied two different methods and as a result we obtained two different values which are in ratio of 1.5 – for example, 26 and 39 points (39/26 = 1.5). The difference between these two cut scores will be equal to 13 points and this is several (more than 4) times larger than the standard measurement error of the test (about 3 points) and can be said to be a statistically significant difference because it is larger than two standard errors (Harvill, 1991, p. 38).
Such a substantial difference between the results obtained by two different methods made Jaeger and many other authors recommend always **to use more than one method** in setting cut scores and the final cut score to be defined only after a careful analysis of all possible data (Jaeger, 1989, p. 500).
This is a very important recommendation and it directly affects the validity and defensibility of the determined cut scores, but, unfortunately, the main question that needs an answer is: *How is it possible for the results from two different methods to be so different if the purpose of all methods is the same – to set the cut score between two*

consecutive and predetermined levels of competence? Actually Glass (1978) asks a similar question and considers this shocking discrepancy to be a stunning blow on the entire practice directed towards setting cut scores (Glass, 1978, p. 249). Hambleton (1978), for his part, does not find anything worrying in the discrepancies from applying the different methods. He explains these differences by the fact that the separate methods have a different scope and purpose and the judgment task itself is different in the different methods (Hambleton, 1978, p. 283). Unfortunately the answer given by Hambleton, although correct, does not solve the problem formulated by Glass. When we are shopping in a supermarket we expect that our bill will be one and the same regardless which cash-desk we choose. The explanation that we were served by different cashiers and this led to a difference in the bill, although the items we bought were the same, would hardly satisfy us. By the same token, students who take a national standardized exam expect that if they had got the same results they would receive the same grades. The explanation that this year the cut scores are higher because we used another method for setting them will hardly satisfy them. According to Zieky (2001) *"If the methods gave different results, people believed that one or possibly both of the results had to be wrong, and there was no way to tell which one was correct."* (Zieky, 2001, p. 35). It should also be added that this is not a question of faith, and is more an instance of a deductive conclusion (if two cut scores derive from the same standard setting context and concern one and the same test, then they have to match or at least to be approximately the same) and "people" should not be blamed for thinking logically.

Nine years after Jaeger's comparative study (Jaeger, 1989), Bontempo together with his colleagues (Bontempo et al., 1998) conducted a meta-analytical study that covered ten comparative studies, four of them included in Jaeger's study. The meta-analysis itself was based on ninety comparisons of two cut scores for one and the same test, obtained by applying two different methods for setting cut scores. One of the major conclusions of this serious and in-depth analysis is that the differences in the cut scores obtained by the application of one and the same method twice are not less than the differences resulting from the application of two different methods (Bontempo et al., 1998, p. 10).

Realizing that the different methods (as well as one and the same method applied twice) can lead to different cut scores, although shocking, is a key moment in the development of the methodology of setting cut scores. It leads to focusing attention on the different factors that influence the process of setting cut scores andon the final results of this process. The direct consequences from this are as follows:

- Greater attention devoted to collecting procedural evidence to support the validity and adequacy of the method used and of the obtained cut scores.
- Greater attention devoted to the issues of the selection and preliminary training of the participants in the judgment task as well as collecting evidence for their competence and their ability to accomplish the given task.
- In-depth analysis of the data related to the internal validity of the chosen method and collect evidence that supports its stability, i.e. if – and to what extent – the results of replicating the method with a different group of judges or different subtests will match.
- Increasingly stronger linking of the judgment to the empirical data and considering this linkage a major argument for the external validity of the method and cut scores that

were set using it.
- Creating a system of criteria for assessing the quality of the methods which will promote a more proper choice of the optimal method in a concrete test situation.

In other words, the main change in the half-century history of the scientific methodology of setting cut scores is switching research from searching for the "true" cut score, which does not exist, towards *"refining and elaborating the systems of rules for deriving and applying judgment"* and improving the argumentation supporting their adequacy and rationality (Cizek, 1993, p. 103).

The hope, or rather the faith, that there is some "true" cut score distinguishing two subsequent levels of competence, is typical of the early stage of development of the methodology of standard setting. The researchers' efforts in this period were towards the development of a method that would help them find this "true" cut score. The reason for this delusion is the parallel with classical test theory, in which the term "true/real value" is one of the main concepts. Starting with Glass, an entire cohort of leading specialists in the domain (Jaeger, 1989; Cizek, 1993; Kane, 1994; Popham, 1997; Hansche, 1998; Reckase, 2000; Zieky, 2001; Linn, 2003) have been trying to refute this myth until at the end a general agreement was reached to the effect that *"... cutscores are constructed, not found. That is, there is no "true" cutscore that researchers could find if only they had unlimited funding and time and could run a theoretically perfect study"* (Zieky, 2001, p. 45), or as Kane (1994) stated in an aphoristic way in the area of standard settings *"There is no gold standard. There is not even a silver standard."* (Kane, 1994, p. 448).

Foreign language testing in Europe is trying to refute this dictum, if not for the cut scores themselves, at least for standard setting. As was already mentioned, the role of the "gold" standard in foreign language testing in Europe is played by the CEFR, whose main goal is to provide *"a common basis for the explicit description of objectives, content and methods..."* to *"enhance the transparency of courses, syllabuses and qualifications, thus promoting international co-operation in the field of modern languages. The provision of objective criteria for describing language proficiency will facilitate the mutual recognition of qualifications gained in different learning contexts, and accordingly will aid European mobility."* (Council of Europe, 2006, p. 1).

Only six years after the publication (2001) of the CEFR (Council of Europe, 2001) it has already reached its goal, becoming the gold standard in the area of foreign language instruction and testing in Europe. There are over 77,000 web-pages where the phrase "Common European Framework of Reference for Languages" can be encountered Indeed, in the Bulgarian Internet-space alone, the phrase "European framework for languages" can be found over 800 times (as of May 2nd, 2007).

According to a survey conducted by the Council of Europe encompassing over 100 European institutions (n = 111) involved in foreign language instruction, the CEFR is very well-known and widely used in the area of testing, assessment and language-proficiency certification (Council of Europe, 2005, pp. 3-4).

This is confirmed also by broader research conducted by Bradshaw and Kirkup (Bradshaw & Kirkup, 2006) focused on internationally recognised certificates in Europe and it includes 273 certificates for 27 languages provided by 48 different institutions. Every year, more than five million Europeans (*n* = 5,336,264) attempt to obtain certificates which report

using the scales for language competence defined in the CEFR.
That is why it can be unequivocally argued that the CEFR played the role of an alarm-clock which announced the end of the age of innocence (according to Zieky's classification) and set the beginning of the age of awakening in the area of setting cut scores in foreign language testing in Europe. This age quickly shifted into the age of disillusionment when foreign-language testing specialists quickly realized that setting cut scores is a complex and contradictory process in which quick and easy decisions are not only wrong, but have a significant counterproductive, negative effect.
During the age of awakening and disillusionment, several new methods for setting cut scores were developed, some of them the object of the current study. The Council of Europe initiated developing guidelines for linking language tests to the CEFR and setting cut scores (Council of Europe, 2003) as well as its concomitant application (Council of Europe, 2004). The major goal of these publications was to facilitate the process of setting cut scores, limit its random nature and increase its quality. In parallel, a series of national and international projects were oriented not only towards setting cut scores for already existing tests, but also (a) towards widening and validating the CEFR, (b) as well as validating the most widely used methods for setting cut scores and (c) creating banks of benchmarks (illustrative examples) for the different performance levels and language skills.

Situation in Bulgaria

Concerning the situation in Bulgaria, it could be argued that we are still in the age of deep innocence. The fact that the topic of setting cut scores found its place in several textbooks in educational measurement and assessment (Bizhkov, 1992; Stoyanova, 1996a; Bizhkov, 2003 and others) does not take us into the age of realistic acceptance because in setting cut scores in the practice of achievement testing in our country, still the most commonly used methods are: tradition, authority and the Goldilocks method, and these methods, as stated earlier, are typical of the age of innocence.
A search of the electronic catalogues of the National Library "St. St. Kiril and Metodi" and the library of the Sofia University "St. Kliment Ohridski" as well as the Bulgarian Internet-space, for example, managed to identify only three cases of the application of methods for setting cut scores, different from the aforementioned traditional methods, and well-known to the scientific society, namely:

- the Nedelsky method, applied to a test for the specialty "Machine fitter" (Simidchiev, 1996);
- the method of the borderline group that was used for setting cut scores in the school readiness test (Stoyanova, 1996b);
- a modification of the Angoff method, used for setting cut scores in the exams for professional certification of Linux specialists by the Linux Professional Institute (LPI, 2006).

The method of authority is especially typical of the Ministry of Education and Science. For example, on the Ministry's web-page (http://mon.bg/opencms/opencms/ left_menu/ citizenship/), in the section related to the organization and conduct of exams for establishing the level of command of the Bulgarian language in order to obtain Bulgarian citizenship, it is stated that *"The exam is conducted in the form of test which consists of twenty items and the time for the examination is one astronomical hour. The purpose of the*

test is to examine the basic communication abilities of the candidates for Bulgarian citizenship and requires elementary knowledge in phonetics, morphology, lexicology and spelling.”. In addition, two different sample tests are provided and the respective cut score is declared, namely that ***“a test is considered to have been passed if it contains 50% +1 correct responses”*** (Ministry of Education and Science, 2007).

Although it is not stated explicitly, it is implied that if the candidate gave 11 correct responses from a total of 20 items, then he or she has *“elementary knowledge in phonetics, morphology, lexicology and spelling”* and can communicate in Bulgarian. In this case, the fact that sample variants of the test are provided is praiseworthy because it can serve as an orientation for candidates as to the nature and difficulty of the exam. However, what the Ministry of Education and Science did not do was to justify the cut score that had been set – 50% +1 correct responses. A decision like this really needs justification, not only because of its significance, or because professional codes require this, but also because this is even inconsistent with the traditionthat the cut score should be about 70% correct responses. The need to provide reasons is even greater in cases like this, when there is a verbal description of the performance standard and the test results are interpreted in the light of this verbal description (in this case – the applicant has/does not have elementary knowledge in phonetics, morphology, lexicology and spelling and can/cannot communicate in Bulgarian).

Such an authoritarian approach to setting cut scores is apparent not only in the tests from the Ministry of Education and Science, but also in most achievement tests that can be found on the Bulgarian book market as well as in some tests for entrance examinations for higher education institutions. For example, in the program for the entrance exam in philosophy in the Faculty of Philosophy at the Sofia University “St. Kliment Ohridski” in 2007 (http://forum.uni-sofia.bg/filo/display.php?page=kand_s) a table for converting raw scores (max=100) into a six-point marking scheme is presented. Thanks to its “high precision” of 0.25 points in fact it defines **13** (!) cut scores for the total raw score because it breaks it into 14 categories/marks:

2.00 | 3.00 | 3.25 | 3.50 | 3.70 | 4.00 | 4.25 | 4.50 | 4.75 | 5.00 | 5.25 | 5.50 | 5.75 | 6.00

Having in mind that the difference between two cut scores is five points of raw scores (with a maximum raw score of 100 points) and that the standard error of measurement in such a maximum raw score is usually greater than five points, obviously the difference between two cut scores is in the range of one standard error. In other words, two cut scores are in fact indistinguishable from each other.

Test theory defines a so called **separation index**. It shows into how many distinct categories a group of candidates can be divided on the basis of their test score (Wright & Masters, 1982, pp. 105–106; Fisher, 1992; Wright, 1996; Schumacker, 2003). In this concrete case, since it groups the candidates into 14 different categories, this index should be equal to 1. However, the separation index (***G***) is directly linked to the reliability (α) of the measurement instrument $G = \sqrt{\alpha / (1-\alpha)}$ and therefore the reliability of this entrance exam as a measurement instrument must be over 0.995 (see Table 1). In point of fact, with this format of exam, such a coefficient of reliability cannot be reached.

Using such a large number of cut scores shows that the authors of this measurement instrument not only live in the age of complete innocence concerning the methodology of

setting cut scores, but they are all in the same period concerning test theory. What is worrying in this case is that, based on their formulation, decisions will be made that concern the future of thousands of applicants for the university.

Table 1 Separation indices and reliability of the tests

Number of categories	2	3	4	5	6	7	8	9	10	11	12	13	14
Cut scores	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Required reliability: ≥	.800	.900	.941	.962	.973	.980	.985	.987	. 990	.992	.993	.994	.995

A careful analysis of the values in Table 1 shows something else – even the conversion from a raw test score to a six-point marking scheme, that has in fact five separation categories, suggests the need for a very high coefficient of reliability (≥0.962), which is almost impossible to obtain in reality. This is confirmed in the studies of Ercikan and Julian (2002, p. 290), according to which reaching a classification accuracy over 80% (which is desirable) is highly unlikely with more than four levels of separation.

This is why one of the most common recommendations related to setting cut scores is to always avoid them when this is not absolutely necessary, and when they are set, their weight should be minimal (Glass, 1978; Shepard, 1980; Zieky, 2001; Ercikan & Julian, 2002; Linn, 2003 and others).

The problem in following this recommendation in the Bulgarian context is that the traditional six-point marking scheme in our country has five assessment categories (fail, average, good, very good and excellent). The situation is similar with the CEFR where the number of competence levels is even higher (six). The problem with the classification accuracy in the context of foreign-language testing is solved most often by using not only one, but several tests, with different tests with different difficulty (usually covering no more than three levels of competence). In this way, the number of cut scores that have to be set is lower than five (most often one, rarely two or three). Another way of solving the problem of a large number of classification categories is to avoid the transformation of the raw test score into a six-point marking scheme in all cases when it is not necessary. In the case of the university entrance exam this transformation can easily be avoided.

One of the most widespread ways for transforming a raw test score into a six-point marking scheme in our country is applying some formula for linear transformation of the following kind: Mark = 2 + *k.n*, where ***n*** is the score from the test for the particular test taker and ***k*** is a coefficient that depends on the maximum possible test raw score. In the case where the maximum possible raw test score is 100 (as in the entrance exams after grade 7 or maturity exams) the coefficient will be equal to 0.04. This means that the cut score (pass/fail) for the maturity exam is set **mechanically** to 25 points of raw score (3 = 2 + 0.04 . 25) regardless of the subject and the difficulty of the concrete test!

This mechanical approach towards setting cut scores is hardly better than current scientific methods (see the table in Appendix 1), but it has already become state policy. For instance, in all directives for the state content standards for obtaining qualification in certain

professions, which until now (May 5th, 2007) are 66 in total, the marking system requires *"assigning of a formula (scale) for calculation of the mark in a six-point marking scheme"* (http://www.navet.government.bg/navet_vq/doi_dv/Prof-doi.htm).
Moreover, in these directives, a verbal description of the separate performance standards is given (see Table 2), which in most cases coincides almost verbatim with the example description given in the didactical directions of the National Agency for Vocational Education and Training (NAVET, 2004).

Table 2 Performance standards

Mark	In theory	In practice
Fail 2	The trainee does not understand basic terms, is not able to define and does not know fundamental facts, processes, laws and subordinations;	The trainee is not able to perform safely particular operations;
Average 3	The trainee knows fundamental terminology in the profession, is able to reproduce the acquired knowledge;	The trainee performs the assigned practical task in accordance with the instruction given in advance, following the regulations for occupational safety; makes mistakes in his work; does not work with precision using laboratory equipment;
Good 4	The trainee understands and explains subordinations, the natural order of things and principles in the processes, methods and subjects under study; is able to apply the acquired knowledge in known situations;	The trainee performs the assigned practical task independently following instructions given in advance, following the regulations for occupational safety; makes small mistakes in his work; works in a proper way using laboratory equipment;
Very good 5	The trainee applies the acquired knowledge in changing situation; is able to plan the accomplishment of the assigned tasks; selects methods for analysis depending on the subject of study and the conditions to be conducted;	The trainee performs completely independently the assigned practical task following the regulations for occupational safety; gives reasons for the approach and order of the analysis; discovers and analyses faults in his work; works accurately with laboratory equipment;
Excellent 6	The trainee applies the acquired knowledge in new situation, is able to plan and justify the completion of the assigned tasks; selects and gives reasons on the selected methods of analysis and control depending on the subject of the studs and the conditions of implementation; evaluates and interprets the results according to the criterion provided in advance.	The trainee performs completely independently the assigned practical task following the regulations for occupational safety; gives reasons for the approach and the order of the analysis; works accurately and precisely with laboratory equipment; evaluates the results from the analysis and evaluates his own work.

In other words, regardless of what test/exam the future tailors, painters, foresters, hotel-keeper, financiers, programmers or upholsterers will take, if they obtain 25 out of the 100-point raw score then, according to the respective directives, it follows that they *know the fundamental theory in the profession and are able to reproduce the acquired knowledge*. Such a *"magical"* solution to the problem of setting cut scores simply by ignoring it, is not typical even for the age of innocence. Unfortunately, the comparison that this approach calls to mind is related not so much with scientific publications on these problems, but with Alice and the Looking-glass and Terry Pratchett's discworld.
Moreover, the appendices of Directive №3, issued by the Ministry of Education and Science

concerning the "education examination standards" (which are clearly performance standards) (changed and complemented in 2004, see issue 47 of "Official Newspaper" from June 1st, 2004), reveal some variety in the test item formats and the cut scores, which is understandable. Different subjects have their specific characteristics, which have to be taken into account in choosing the test item formats as well as in the system for the evaluation of the results. However, with the changes in, and the supplement to, this directive, that came into force on December 5th, 2006, this variety has been removed and complete unification of the examination programs (performance standards) has been imposed. Moreover, inexplicably, article 3, section 5 (requirements of assessment) has been abolished.

If the main characteristics of the Ministry of Education and Science's initiative for *"overall introduction of tests"* in Bulgarian education can be summarized, the main word is **absence**: absence of transparency, absence of competence, absence of variety (in test format, in performance standards and methods for setting cut scores), as well as absolute ignoring of professional standards (both ours and internationally acknowledged ones) and international experience in this domain. Bearing in mind that one of the meanings of the word *innocence* is lack of knowledge and understanding, this absence is typical of the first age in the categorization made by Zieky (1994). The bad news is that the age of innocence is followed by a painful awakening and subsequent disillusionment.

2 Methods for setting cut scores

2.1 Review of methods

The first method for setting cut scores in criterion-referenced achievement tests was created by Nedelsky in 1954 (see Nedelsky, 1954). Thirty-two years later the number of methods had already reached 38 (Berk, 1986) and it is still increasing. For example, in the process of selecting appropriate methods for setting cut scores for the National Assessment of Educational Progress in USA Reckase (Reckase, 2000a) described and evaluated 14 newly created methods for setting cut scores. During the same year Hambleton and his colleagues (Hambleton et al., 2000) presented 10 other methods for setting cut scores which are appropriate mostly for productive skills.
To date, over 60 have been documented, with a large number of them listed in Appendix 1.

2.1.1 General description of methods

At the very outset, it should be stressed that the list of methods for setting cut scores in criterion-referenced tests in Appendix 1 is not complete. For example, only 13 out of the 38 methods described by Berk (Berk, 1986) are listed. The reason for the exclusion of the remaining 25 methods is that they have not been applied in the last 20 years. In other words, the main criterion in selecting the methods to be included in the list is that they had been used in the last 20 years. The list of methods in the table is sorted in ascending order according to the year of their creation (column 4), and as can be seen, two-thirds of them were created between 1990 and 2005.
The table in Appendix 1 contains 12 columns in total. Here is a short description of each column and of the codes and abbreviations that were used:
Column 1 (№) – the sequential number of each method listed in the table.
Column 2 (**Method**) – the name of the method.
Column 3 (**Source**) – the source from which the method is cited. The full reference list of the separate sources (books or articles) can be found in the References at the end of the book.
Column 4 (**Year**) – the year in which the method was published. As was already mentioned, the table is sorted according to the date in this column.
Column 5 (**Focus**) – the focus of judgment is indicated. The subject of judgment can be the test items (**test items**), the test takers (**test takers**), the test takers' responses to the test items (**item resp.**), the test score (**test score**) or the test profile (**profile**) in such cases where the test score consists of several components. In case the judgment has more than one focus, the major one is indicated and a "+" sign is added.
Column 6 (**Rounds**) – the number of rounds in the judgment process.
Column 7 (**Emp. data**) – indicates whether the judges were provided with concrete empirical data (**yes/no**) concerning the test results.
Column 8 (**Format**) – any limitations in usage of the particular method concerning appropriate item types are indicated. For example, if the method is appropriate only for items with an extended constructed response, this is indicated by the notation "**ext. resp**" in this column. If the method is appropriate only for multiple-choice items, the notation is

"**m. c. resp.**". In cases when no such limitation applies the entry will be "**no**".
Column 9 (**Scoring**) – limitations of the method concerning the scoring of the test item responses: "**dich**." indicates methods appropriate only for dichotomous scoring (correct/incorrect) and "**polyt.**" for methods appropriate when the scoring of the items is polytomous. If such limitations do not exist, the entry in this column is "**no**".
Column 10 (**Model**) – indicates limitations (if present) concerning the test model. When the method is appropriate only for tests constructed and analyzed by Item Response Theory, then the entry is "**IRT**". If the method is applicable regardless of the model, the entry in this column is "**no**".
Column 11 (**Emp. data**) – indicates (**yes/no**) whether, in the final setting of the cut scores, concrete empirical data is taken into account besides the judgments.
Column 12 (**Stat. meth.**) – indicates the statistical methods used in setting the cut scores. Due to the limited space in the table cells, the following abbreviations were used:

- **descr. stat.** – descriptive statistics
- **lin. regr.** – linear regression
- **nonl. regr.** – nonlinear regression
- **mult. regr.** – multiple regression
- **log. regr.** – logistic regression
- **clust. an.** – cluster analysis
- **prob. theor.** – probability theory
- **IRT** – Item Response Theory
- **math. opt**. – mathematical optimization
- **fuzzy sets** – fuzzy set theory

The "+" sign in this column is used in cases where the descriptive statistics is the main, although not the only, statistical method used in setting the cut scores.

2.1.2 Classification schemes

Although incomplete, the list of available methods is quite long and obviously needs some systematization.

Meskauskas, Hambleton and Berk

One of the earliest classifications of standard setting methods is attributed to Meskauskas (1976) who divided methods into two main groups: *state models* and *continuum models*. The difference between the two kinds of models lies in the kind of variable that the competence is expressed in – a discrete variable with clearly defined levels (state models) or in terms of a continuous quantity (continuum models). In the modern concept of the development of different competences they are conceived of as continuous variables. That is why Jaeger (1989, p. 493) recommends that the choice of the method for setting cut scores in real situations should be limited to those belonging to the group of continuum models. In practice, all of the 62 methods listed belong to the category of "continuum models", which in turn shows that this classification scheme is inappropriate for the present situation and has mainly historical value.
Several years later Hambleton (1980, pp. 103-107) suggested a different classification scheme, which divides the methods into three major groups: *judgmental, empirical models*

and combination models. The main criticism of this classification scheme is that these names mask the fact that all methods contain a subjective component. This, according to Jaeger (1989, p. 493), could lead to confusion and erroneous interpretation.
Drawing on the classifications by Meskauskas and Hambleton, Berk (1986) suggested a more complex classification scheme, which includes three levels:
The first level is related to the nature of the measured competence and in practice coincides with the classification made by Meskauskas, dividing the methods into state and continuum models.
The second level of the classification scheme groups the methods depending on the relative amount of subjectivity in setting cut scores. Berk (1986, p. 139) divided methods into three main groups:

- **judgmental** – the cut scores are set only on the basis of judgment;
- **judgmental-empirical** – the cut scores are set mainly on the basis of judgment which, however, takes into consideration empirical data;
- **empirical-judgmental** – empirical data are a decisive factor, but judgment is also taken into account.

As can be seen from this classification, a subjective element is always present, but its share varies. Typical examples of judgmental methods are the classical ones of Nedelsky (first in the table), Angoff (2 and 3) and Ebel (4). An example of the judgmental-empirical methods is the one by Jaeger (6). Two well-known examples of the empirical-judgmental groups of methods are – the Borderline Group method (7) and the Contrasting Groups method (8).
The third level refers only to empirical-judgment methods, dividing them into two groups depending on whether they are intended for setting cut scores or only for the *adjustment* of cut scores that have already been set. Since the table in Appendix 1 includes only methods for setting cut scores, and not for their adjustment, all of them come under the first group of the classification scheme. Berk described eight examples of models for adjusting already set cut scores, among which are, for example, Emrick's (1971) and Macready and Dayton's (1977) models.

Jaeger and Reckase
However, the most widely used classification of methods for setting cut scores is the one by Jaeger (1989, p. 493). According to this classification, methods are divided into two main groups depending on the focus of the judgment as follows:

- **Test-centered** methods, in which the subject of the judgment is the elements of the test (mainly test-items);
- **Examinee-centered** methods, in which the subject of the judgment are the examinees, i.e. those whom the test-items are intended for.

If this classification is applied to the methods described in the table in Appendix 1, only 5% of them belong to the group of examinee-centered (7, 8 and 9) while all the rest are in the group of test-centered methods. There are several reasons that could explain this lack of balance:

a test-centered methods are preferred from a practical point of view, since they require fewer resources and their organization is easier;

b In spite of the lack of evidence, traditionally, the test-centered methods are assumed to be more reliable and valid;
c the increasing use of test formats that are appropriate for measuring productive skills *(performance assessment)* in the last 15 or 20 years led to the creation of new methods for setting performance standards, appropriate specifically for this test format. It is typical of these methods that the focus of the judgment is not the test itself but the test takers' responses to the test-items. Since a significant share of the methods for setting cut scores are of this kind, some authors (Haertel & Lorie, 2004, p. 80) list these methods in a separate, new (additional) category: **performance-centered** methods.

Although it is still widespread, Jaeger's classification is outdated because it is not able to encompass well enough the entire variety of current methods for setting cut scores. That is why in the last few years new classification schemes have appeared.

For example Reckase (2000a, pp. 46-49) uses a classification scheme consisting of three levels, depending on:

a the level of the complexity of the task assigned to the judges;
b the quantity and type of information and/or empirical data provided to the judges;
c the level of complexity of the procedures used for summarizing the results of the judgments and the final setting of the cut scores.

Although logical, this classification scheme has not been widely adopted. The probable reason for this is due to the fact that each of the three levels is considered a continuous variable and not a discrete one.

Hambleton, Jaeger and Plake

The most promising modern classification scheme is the one suggested by Hambleton, Jaeger and Plake (Hambleton et al., 2000, pp. 356-357). This scheme contains six levels related to the focus, assignment and process of the judgment, the selection of judges, validation of the cut scores and the characteristics of the instrument. From all six levels, only the first two have been operationalised so far – the ones concerning the focus of the judgment and the assignment given to the judges. The first level is actually an extension of Jaeger's classification and includes four major categories, which can be labelled as follows:

- **Test-centered** – the subject of judgment is the test-items (task material);
- **Examinee-centered** – the subject of judgment is the examinees, i.e. the ones that give response to the test-items;
- **Performance-centered** – the subject of judgment is the responses of the examinees to the concrete test-items (work products);
- **Result-centered** – the subject of judgment is the tests scores (and/or their frequency distribution) or the test profile (in case the test is comprised of several components – sub-tests).

As can be seen, the first two categories of this classification scheme match the categories of Jaeger's classification, while the other two are consistent with the modern tendencies in the development of setting cut scores.

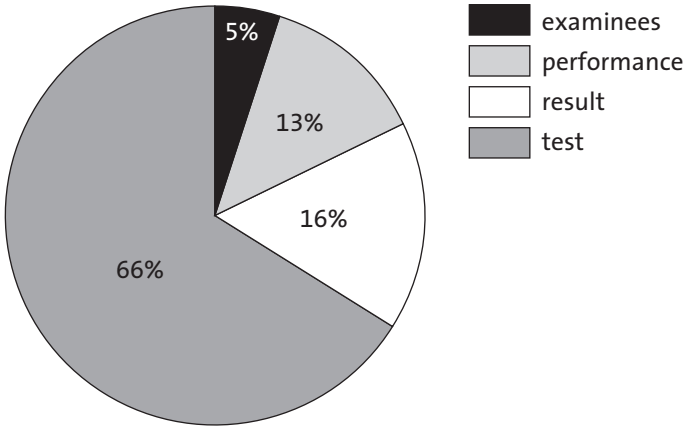


Figure 2 Distribution of the methods depending on the focus of judgment

Test-centered methods

Figure 2 represents the frequency distribution of the methods for setting cut scores listed in Appendix 1, according to the above classification. The figure shows that the majority of the methods (two-thirds) are **test-centered**. This is easy to understand bearing in mind that, on one hand, the multiple-choice tests for which these methods are most appropriate are still the predominant test format in achievement testing. In addition, as already mentioned, from a practical point of view, test-centered methods are preferable because compared to other types they require less advance preparation, less time and fewer resources. A considerable portion of these methods (68%) can be applied a priori without even administering the test, which is one of their main advantages.

Another advantage of these methods is that, in applying them, significantly more judges can be used compared to the examinee-centered tests, which is an advantage for obtaining a higher level of reliability in the judgment task.

On the other hand, however, all test-centered methods require the judges to judge the difficulty of the test items in one way or another. This is usually done by judging the probability for a correct response in some defined target group of examinees or by classifying the test items into several predefined groups depending on what level of competence is required in order to give a correct response to them. Judges' ability to correctly classify the items by their difficulty is the subject of numerous studies (van der Linden, 1982; Smith & Smith, 1988; Livingston, 1991; DeMauro & Powers,1993; Impara & Plake, 1998; Goodwin, 1999; Chang, 1999; Plake & Impara, 2001; Chang, L. et al., 2004 and others). The main conclusion drawn from these studies is that, as a whole, judges cannot judge the item difficulty adequately, especially if they did not go through preliminary, tailor-made training. This problem is considered the main disadvantage of test-centered methods for setting cut scores.

Examinee-centered methods

The three methods that fall within the category of **examinee-centered** methods (7, 8 and 10 in the table of Appendix 1) were created in the early stage of the development of standard setting methodology. Their main advantage is that the cut scores derived using these methods usually have a high degree of stability and adequacy. From a practical point of view, however, their application meets with several difficulties. The other disadvantage they have is that the number of the judges is usually limited to one for each examinee. This is inevitable due to the nature of each examinee's level of competence. That is why they are being increasingly replaced by the performance-centered methods, although in particular the contrasting groups method is still applied frequently.

Performance-centered and results-centered methods

The **performance-centered** methods are also basically examinee-oriented, but with a narrowed-down focus. This is also confirmed by the name of one of the most widespread methods: *Generalized Examinee-Centered method*, (№ 47 in the table of Appendix 1). However, while in examinee-centered methods, the subject of judgment is the overall behaviour of the examinee **before testing**, in respect of the indicators of the measured characteristic in performance-centered methods the subject of judgment is only the products of examinees' activity **during the testing**. This focus overcomes the main disadvantage of the examinee-centered methods, namely the limited number of judges who can judge an individual examinee. In performance-centered methods, such a limitation is not present since the product, i.e. the examinees' responses, can be given for judgment to an unlimited number of raters.

A significant point that should be noted is that while in Berk's (1986) list, which includes 38 methods in total, performance-centered methods are missing, in the current list (Appendix 1) their relative share is already 13%. This change among the different kinds of methods shows the strong link between theory and practice in achievement testing on one hand and standard setting methodology on the other.

Concerning **result-centered** methods, although they go beyond the performance-centered methods, they are relatively less frequently used. They are mostly used in measuring complex skills, when the test consists of several, differentiated parts or as a concomitant, subsidiary method, used for comparative purposes and for the validation of the cut scores obtained with the help of some other primary method.

2.1.3 Main characteristics of methods for setting cut scores

Figure 3 is a summary of the main characteristics of the current methods for setting cut scores. The figure shows that the percent of methods in which the judgment is implemented either in one or in more than one rounds is slightly less (53%) than for methods requiring a single judgment (47% to 53%). Multiple judgments is more typical of the test-centered (54%) and performance-centered (50%) methods, while in result-centered methods more than one round of judgment is typical only of 30% of the cases, and in the examinee-centered method it is always a single judgment. The bigger relative share of methods employing multiple rounds of judgments in the first two categories of methods is due to the fact that, in these methods, judges are provided with empirical data.

The main goal is to achieve maximum consistency between the judgment and the empirical data and in this way to increase the validity and adequacy of the cut scores. Unfortunately, as can be seen from Figure 3, for the majority of the methods (68%), the judges still do not have any empirical data when making their judgment. While this can be justified in the examinee-centered methods because of their specific conditions, there is no such justification for the rest of the methods. The only justification in this case could be the time factor:

- Time needed for preparing the empirical data;
- Time needed for training the judges in interpreting the data;
- Time needed for ascertaining the concordance between the initial judgment and the empirical data.

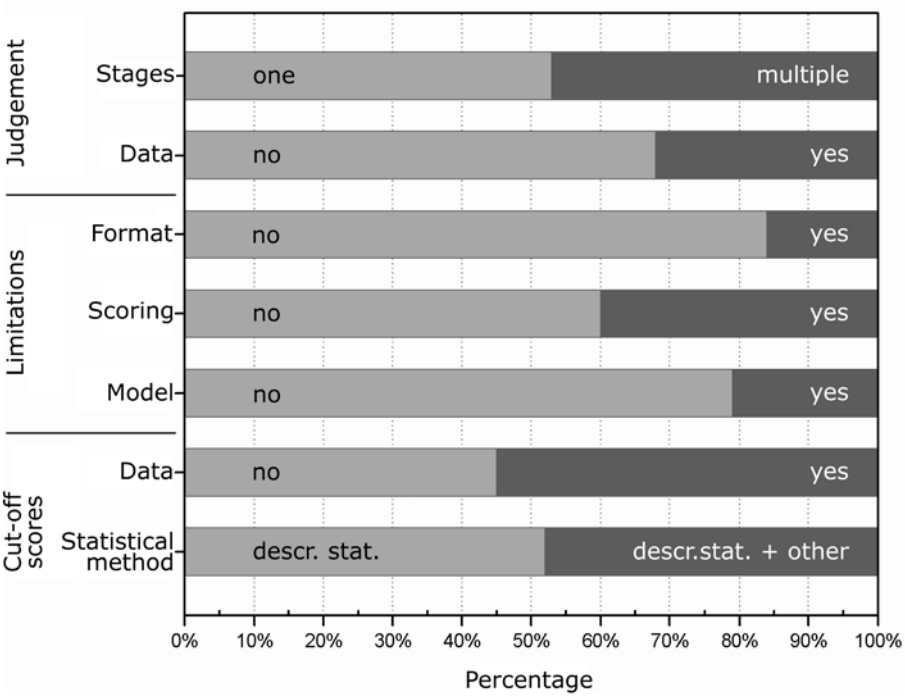


Figure 3 Main characteristics of the methods for setting cut scores

In methods employing a single judgment, the judges have access to the data in only 15% of cases, while empirical data is provided in the majority of the cases in methods employing multiple judgments (52%). The difference in these percentages is statistically different (p = 0.002), which is also confirmed by the corresponding error-bars (Figure 4).

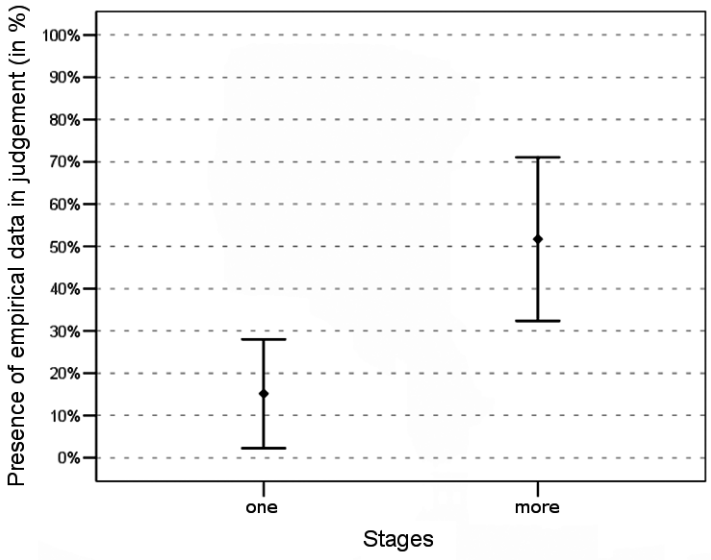


Figure 4 Empirical data in judgment depending on the number of rounds (stages)

Concerning limitations, there are no limitations in the majority of methods, either for the format (84%) or the judgment (60%) or the model that has been used (79%).
As a whole, in the examinee-centered methods, there are no limitations for the format, judgment or the model (see Figure 5). In the other three classification categories, the most limitations are associated with test-centered methods (61%). This is logical since in these methods the items and the item format are the subject of judgment, and the type of judgment and analysis also inevitably influence the major characteristics and eventually pose limits on the particular method for setting cut scores.

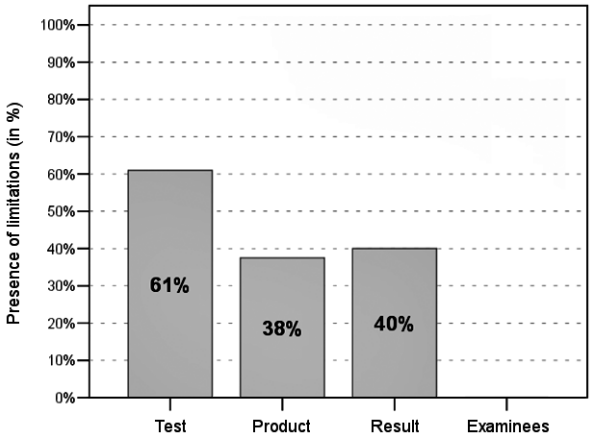


Figure 5 Limitations in standard setting depending on the judgment focus

The failure to provide data, which is the case in many the methods for setting cut scores, is justifiable only if empirical data are taken into account at least in the final setting of the cut scores. This is feasible if the subjective judgments are pooled with the objective results. Unfortunately, the results from the analysis (Figure 3) show that in almost half of the methods (45%) such aggregation is missing.
This finding contradicts one of the main recommendations in the literature on standard setting, according to which the integration of the results from the judgments with the empirical data is the only guarantee for the adequacy of the cut scores (Livingston & Zieky, 1982; Jaeger, 1990; Norcini, 1994; Pellegrino et al, 1999; Zieky, 2001; Linn, 2003 and others). In other words, pooling both kinds of information (subjective judgment and normative data) *"in setting a cut score can help avoid the establishment of unreasonably high or low values"* (Zieky, 2001, p. 38).

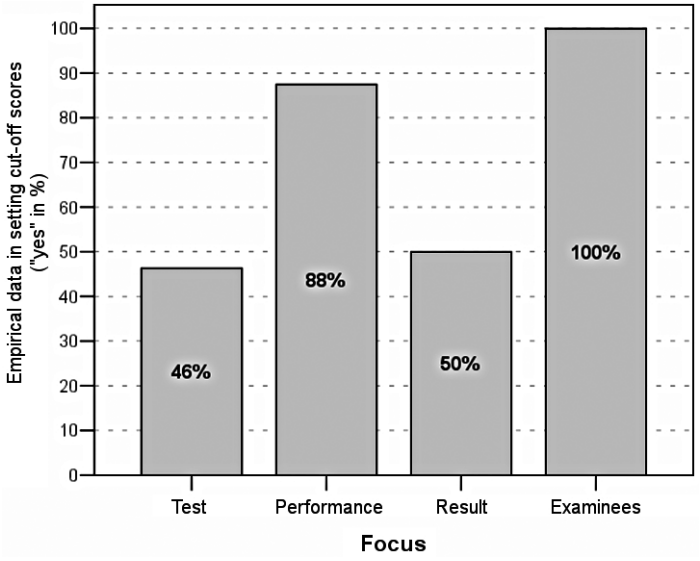


Figure 6 Empirical data in setting cut scores depending on the focus of judgment

As can be seen from Figure 6, linking the final cut scores with the empirical data depends to a large extent on the kind of methods used. For example, in examinee- and performance-centered methods, due to their specific characteristics, the percentage of methods in which the cut scores are set utilising empirical data is over 90%[3]. It is not accidental that, according to Berk's (1986) classification, the examinee-centered methods are in the category of the empirical-judgmental ones.
Concerning the other two groups of methods, which focus on the test items (test-centered)

3 88% for the performance-centered and 100% for the examinee-centered methods or an average of 91% for the two groups.

and the test score, the majority of them are completely judgmental. The setting of the cut scores in these methods is based only on the judgments[4] without taking into account empirical data at all.
The statistical methods used for summarizing the results from the judgments and the final setting of the cut scores are limited to calculating only descriptive statistics for the majority of methods (52%) (most often the measures of central tendency – mean and median).
The other most often used statistical methods are regression analysis and Item Response Theory (IRT).
As can be seen from the pie-charts in Figure 7, there is a tendency for using more complex statistical methods in the newer methods for setting cut scores. At least two reasons for this tendency can be pointed out:

a with the development of information technology, the access to the corresponding software for statistical analysis of the data has become easier;
b the aspiration to increase the reliability and validity of the cut scores imposes higher and higher requirements on the statistical methods used for their setting as well. This, unfortunately, does not mean that the more complex the statistical methods used the more precise and adequate the corresponding cut scores will be!

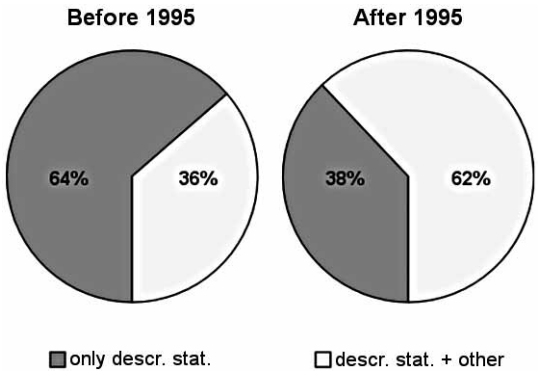


Figure 7 Tendencies in the use of applied statistical methods

Summary
From the above short review of the existing methods for setting cut scores in criterion-referenced achievement testing the main characteristics of these methods can be summed up as follows:

- In the rich arsenal of methods, the dominant are the test-centered ones (66% of the total number).
- In the major part of the methods (68%), the judges do not have at their disposal

4 46% for the test-centered and 50% for the examinee-centered methods or an average of 47% for the two groups.

concrete empirical data to base their judgments on. This contradicts a key recommendation and casts doubts on the validity of the cut scores that have been set.

- The number of the separate rounds of the judgment process varies between one and four and in most cases (53%) is limited to one round only, mostly because of practical reasons. Multiple judgments is more typical of the test- and performance-centered methods and most often this is related to providing empirical data and discussion at rounds II, III or I.
- For almost half of the methods (48%), there are no limitations concerning the test item format, the method of scoring their responses or the models used in the analysis of the test results. The presence of limitations is typical of test-centered methods, while all examinee-centered methods are without limitations.
- A significant part of the methods (45%) can be defined as completely subjective because empirical data are neither provided to the judges nor used in the final setting of cut scores. This casts serious doubt on the adequacy of the cut scores based on these methods and requires external validation. Setting cut scores only on the basis of judgment is most typical of the test-centered methods (in 61% of them). Methods of this kind are most frequently used in practice and – what is more important – as the only method for setting cut scores. That is why the question of the validity of the cut scores that have already been set is a question of great topical interest.
- In the majority of the methods (52%), the statistical methods used in setting the final cut scores are limited only to descriptive statistics. During the last few years, however, a tendency to use more complex statistical methods is evident, mostly due to the increasingly wider use of IRT in large-scale projects.

2.2 An illustrative example

The main purpose of the present study is to describe six methods developed by the author for setting cut scores in criterion-referenced tests and to provide arguments (both theoretical and empirical) in support of their validity and, based on a comparative empirical analysis, to determine the most effective among them for mass use.

Irrespective of what the method is, for its concrete and detailed representation, a concrete measurement instrument (achievement test) is needed. Specific methods for setting cut scores should be applied to this instrument.

On the other hand, since each method for setting cut scores is more or less based on judgment, a set of judgments from several judges who are judges in the corresponding domain is required as well.

Moreover, the cut scores are related to and predefined by the corresponding performance standards, and they are actually their operationalization (Kane, 1994, p. 426). That is why some concrete performance standards are also necessary. On the basis of them the corresponding cut scores will be set.

Due to the fact that the cut scores are to a large extent predetermined by (a) the specific characteristics of the particular measurement instrument, (b) the characteristics of the judgments and the characteristics of the participants who make the judgments as well as (c) the performance standards, it is highly desirable to discount the influence of these

factors if we wish to be able to carry out a concrete comparison of the different methods. This is feasible only if the different methods are applied to the same test with the participation of the same judges and in using the same performance standards.
The following description of the methods and their comparative analysis will first describe how these three conditions of equivalence (instrument, judges and standards) were met.

2.2.1 Instrument

As an illustrative example of an achievement test to which the six new methods for setting cut scores will be applied, a set of 27 test items for reading comprehension will be used. This set of test items is part of a computerized bank (BITE-IT: Bergen Interactive Testing of English), which is used in annual external assessment in English (reading comprehension) in grades 4 through 11 in Norway.
The parameters of the separate test items are established on the basis of pilot studies in which an *incomplete linked test design* was used. Each item was included in at least two test booklets and was answered by at least 450 examinees. The computerized bank is based on IRT using a one-parameter logistic model – OPLM (Verhelst et al, 1995). The scales used for reporting the results for the separate grade levels (4, 7, 10 and 11) are different, but linked to each other. This allows comparability of the results both across the years in which the tests were administered and also across the grades.
The external assessment itself is computerized and is conducted on-line, and students and their teachers receive the results automatically right after the completion of the test.
The major reasons for choosing this concrete set of test items as an illustrative example are as follows:
First, this is a criterion-referenced achievement test and the six methods for setting cut scores are aimed exactly at instruments of this kind.
Second, the test is aimed at measuring the level of reading comprehension in education in English at level A2/B1/B2 according to the CEFR. In other words, the format of the instrument corresponds to the frequently used format in reading and listening comprehension tests used in foreign language testing. This is of great importance since all of the methods that will be described are developed for the needs of foreign language testing and, moreover, particularly for cases of linking the tests to the CEFR.
Third, the test is relatively short – it consists of only 27 test items, which makes it especially suitable for illustrative purposes. At the same time, however, despite the short length, its reliability is high enough (0.91) to provide the necessary classification precision in transforming the test score into at least three classification categories (Table 1). In other words, the psychometric properties of the test allow at least two cut scores to be set. This is an important criterion in selecting the illustrative material because all methods that will be presented allow the setting of more than one cut score and it would be good if this property is demonstrated in practice.
Fourth, Item Response Theory is a very powerful theoretical tool, but for its successful application on a given test, the test itself has to satisfy a considerable set of preliminary requirements which are often hard to meet. The chosen test, however, meets these requirements. More specifically, for this test, the empirical data fit the one-parameter Rasch model. As already mentioned, for the computerized bank to which this set of items

belongs, a more complex probability model is used (one-parameter logistic model - OPLM), but its application in this illustrative example would lead to an additional complexity of the presentation. That is why, for illustrative purposes, the simpler, and at the same time the most perfect, model of IRT was chosen – the Rasch model (Stoyanova, 1996c).
Fifth, the test items for this project were developed to reflect the performance standards for language competency (reading comprehension) of the CEFR. In other words, predefined performance standards exist, and this is an important and necessary prerequisite in setting the corresponding cut scores.

2.2.1.1 Test description

The stimulus material for this set of *items* consists of 17 short texts (between 19 and 99 words each) and 3 longer texts with lengths of 154, 156 and 165 words. The number of items for a given text varies between one and five and usually more items are associated with the longer texts.
A large part of the items (10) require a selected response and the item options are words or phrases (8) or images (2). The number of options for this type of items is three or four.
The other item format, frequently used in computerized testing, is marking text (or separate words) and most items (11) are of this kind.
In addition, there are five matching items and one that require moving an image in accordance with given instructions.
The scoring is dichotomous and is done automatically based on a predefined key for the correct answers.
Each item was piloted using a sample of at least 462 students, but since the pilot study uses an incomplete linked test design, the different items were responded to by different examinees. That is why, for simplifying the current presentation, only the results from the sample of examinees answering all the items (27) will be presented. This results in a sample of 250 examinees.
The pilot study was conducted in the spring of 2006 and comprised 3742 seventh-grade students in Norway whose suggested average level of language competence (English – reading comprehension) was A2/B1. The total number of items piloted was 552.
The test itself (texts, test items and their scoring schemes) is not published due to the need to maintain test security, but since the concrete content of the texts and the corresponding items is not directly relevant to the current discussion, this non-disclosure is perfectly acceptable. A demonstration version of the test that illustrates the different formats of the test items and the level of their difficulty can be found at the following web-address: http://bite.uib.no/npweb.

2.2.1.2 Psychometric characteristics of the test

Table 3 summarizes the psychometric properties of the test; the psychometric properties of the individual items are presented in the table in Appendix 2.
Due to the relatively small sample (n = 250), the number of examinees with the same raw test score is not large, either (between 0 and 20). For the analysis of the behaviour of the item and examinees that are located around the corresponding cut score, it is better to work with a larger sample. That is why, for this illustrative example, a data set of

5000 examinees was generated with the aid of a computer simulation model.
The method of simulation modelling is frequently used in scientific research when the purpose is the in-depth analysis of a given process (in this case – setting cut scores) and the analysis of the effect of different (controlled) initial conditions and parameters.
The basic, predefined parameters for generating the simulated data on were as follows:

- Length of the test – 27 items;
- Item scoring – dichotomous (correct -1; incorrect -0);
- Item difficulty – the same as the estimated item parameters in the real test, when using the Rasch model.[5]
- Discrimination index – approximately the same, equal to the mean item discrimination index for the real sample (+0.53);
- Sample size – 5000 examinees;
- Frequency distribution of the latent proficiency – normal, with a mean of 0 and a standard deviation of 1.

The computer generation of the data using the aforementioned parameters was done with the program OPSIM, which is a module of the OPLM software for IRT (Verhelst et al, 1995, pp. 112-115).
As can be seen from the results in Table 3, the psychometric characteristics for the real and simulated data are not substantially different. The only more significant difference is that the width of the interval in which the discrimination index varies is smaller for the simulated data (0.20 vs. 0.37). The reason for this is in the predefined condition for an approximately equal discrimination index. The necessity for this condition was imposed by the desire for matching the simulated data with the chosen one-parameter model, which suggests an equal discrimination index for the test items.
Such a correspondence, especially in such a large sample size (n = 5000), is hardly feasible in practice, but as it can be seen in Table 3, it is achieved for the simulated data. Moreover, the correspondence between the model and the data is better for the simulated data although the sample is 20 times larger.

5 Editors' note: We have removed a reference to a Z-scale in the original, as that appears to have been an inadvertent error by the author. This also applies to the last point in the list.

Table 3 Psychometric characteristics of the test

Parameters		**Sample (real data)**	**Sample (simulated data)**
Number of examinees		250	5000
Number of items		27	27
Difficulty	Minimum	16%	17%
	Mean	50%	50%
	Maximum	88%	87%
Discrimination index	Minimum	0.33	0.39
	Mean	0.53	0.54
	Maximum	0.70	0.59
Test score	Maximum	27	27
	Mean	13.56	13.52
	Standard deviation (*SD*)	6.66	6.71
Reliability (α)		0.91	0.91
Standard error (*SEM*)		2.00	2.01
Correspondence between the model and the data	Items (*p*)	$p > 0.02$	$p > 0.01$
	Test (*p*)	$p = 0.053$	$p = 0.084$

As already mentioned, this correspondence predefined the choice of this test for illustrative purposes. The reason is that in the Rasch model, all items have the same weight in setting the total test score (which is not the case in other IRT models). This attractive characteristic of the model will facilitate the description of the methods for setting cut scores.
The other important advantage of all IRT models is that they allow the item difficulty and the measured construct (in this case – reading comprehension) to be represented on the same measurement scale (in this case – standardized Z-scale, fixed in terms of the item difficulty: $Z_{mean} = 0$; $SD_Z = 1$).
Figure 8 demonstrates this advantage, representing the frequency distributions of the item difficulty and the reading comprehension in English of the students tested with this instrument and in the simulated data. The horizontal axis in the figure represents the underlying scale. As can be seen in this graph, as a whole the mean difficulty of the items matches the mean level of the measured construct both for the real and for the simulated data. Moreover, there is almost a perfect match of the frequency distributions in both types of data (real and simulated) and there is a good coverage by the items of the interval (-2.8; +2.4) in which 91% of the real examinees and 88% of the simulated data fall.
Figure 8. Frequency distributions for the item difficulty and the ability for reading comprehension of the examinees

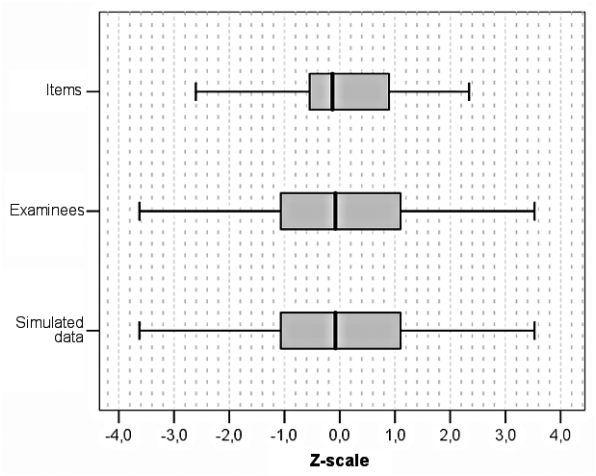


Figure 8 Frequency distributions for the item difficulty and the ability for reading comprehension of the examinees

Although item difficulty varies widely (between16% and 88% per cent correct), this does not affect their discrimination index, which is high enough both for the least (> +0.32) and most (> +0.49) difficult item. The high discrimination index of the test items is the reason for this test having high (0.91) reliability in spite of its small length (27 items).
The high reliability, both for the real and simulated data, allows test score transformation into more than two classification categories with a sufficiently high classification precision (see Table 1 in Ch. 1). This, in turn, will make it possible to set more than one cut score, which is an important feature of all the methods to be discussed in this thesis.
Appendix 3 contains the table for transforming the raw test score (number of correctly answered items) into a value on the underlying scale. Since the standard error of the measurement *(SEM)* represents the averaged value of the error across the different points of the test score scale, in the transformation table in Appendix 3, the error for each value is provided (in brackets) – the so called *conditional standard error of measurement – CSEM*.
To find the conditional standard error for the raw test score, Keats' modification of Lord's binomial method (Feldt et al, 1985, pp. 353-354) was used, and for the underlying scale, the classical approach, based on the information function of the test (Lord, 1980; Hambleton et al, 1991) was used.

2.2.1.3 Performance standards

The methods for setting cut scores, which are the subject of the present study, were developed for the needs of foreign language testing and, more precisely, for interpreting test results in the context of the levels of competence defined in the CEFR (Council of Europe, 2001).
That is why, in setting the cut scores for the illustrative example, the performance standards for reading comprehension described in the CEFR will be used.

According to the CEFR, language competence is divided into six main levels as follows (Figure 9):

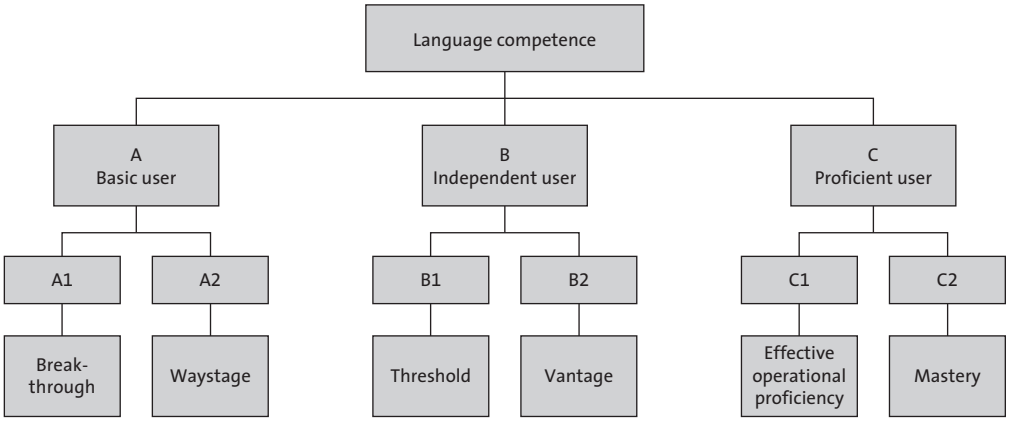


Figure 9 Levels of language competence

The verbal description of the performance standards for each of the six levels is given through a set of proficiency scales. The CEFR contains a total of 57 scales. The large number of the scales is required by the great variety of language skills, communicative activities and strategies that are applied in communication and in the accomplishment of different communication tasks.

For example, performance standards for the separate levels of competence in reading comprehension exist for nine assessment scales (Council of Europe, 2001), namely:

- Global scale (p. 24);
- Self-assessment grid (pp. 26-27);
- Overall reading comprehension (p. 69);
- Reading correspondence (p. 69);
- Reading for orientation (p. 70);
- Reading for information and argument (p. 70);
- Reading instructions (p. 71);
- Watching TV and film (p. 93);
- Identifying cues and inferring (Spoken & Written) (p. 72).

The total number of the descriptors that characterize the different levels of competence in reading comprehension in these assessment scales is 56. For the separate levels they are, respectively, 7 for A1, 12 for A2, 13 for B1, 12 for B2, 5 for C1 and 4 for C2.

A complete list of the performance standards in reading comprehension for the respective levels of competence is provided in Appendix 4. The performance standards are presented in their English version (CEFR, 2001) for two reasons:

First, the English version was used for setting the cut scores of the test in the illustrative example.

Second, at present there is no research or evidence supporting the equivalence of the versions of CEFR in English and Bulgarian.

In the description of the instrument it has already been mentioned that the expected level of competence of the target group (seventh grade students from Norway) was A2/B1/B2. This logically brings up the question of limiting the list with performance standards (Appendix 4) only to those that correspond to these three levels. Such a limitation is, however, not desirable because, due to different reasons, the test might contain test items that can be answered correctly by examinees who are on a lower (A1) or a higher (C1 or C2) level of competence. That is why the judgment in this type of methods for setting cut scores is usually conducted by providing judges with the performance standards that describe the entire continuum of the measured characteristic. The judges are those who, through their evaluation activity, pose the limitations on the range of levels for the respective competence.

A distinctive characteristic of the performance standards in the CEFR is that, for each level of competence, they provide a detailed description of what each examinee who is on this level knows and is able to do and in what circumstances. For instance, according to the CEFR, a person who is at the basic level of language proficiency (A1) *"Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help."* (Council of Europe, 2001, p. 24).

Another significant distinctive characteristic of this system of performance standards is that even the examinees who are at the lowest level of competence have some (although limited) knowledge and skills.

These two distinctive characteristics of the performance standards defined in the CEFR determine the area of application of the methods for setting cut scores, which are the subject of the current research. These methods can be applied only when a system of performance standards is created in advance and the standards provide a detailed description of what an examinee should know and be able to do at a certain level.

Situation in Bulgaria

These methods for setting cut scores **cannot** be used for transforming the test score into the current six-point marking scheme in Bulgaria due to two main reasons:

First, there is no system of performance standards (PLDs) which describes in a detailed way what, in concrete terms, each student must know and be able to do for a particular level of competence (fail, average, good, very good and excellent), neither for a particular grade nor for a particular subject. The so called "qualitative" and "quantitative" indices for assessment in the current Directive № 3 of the Ministry of Education and Science (2005) are limited only to an enumeration of the five levels of competence and their labels: excellent 6, very good 5, good 4, average 3, fail 2. In the directive, there is not even a hint of a description of what a given student should know and be able to do to obtain the respective mark. Concerning the criteria for assessment in the National Examination Programs, which the directive refers to, they are not given by levels, but with a specified weight, and the main purpose is for the sum of the weights of all criteria to be equal to a number fixed in advance – most often 60 (or 100). In assessing a certain examinee, his/her sum of test

points is transformed mechanically, with the aid of a "magic formula", into the respective mark which, after rounding, sets the level of competence for this examinee (excellent 6, very good 5, good 4, average 3 or fail 2).
Second, even if there are some attempts at setting performance standards (e.g. http://rio-varna.hit.bg/R-OcenqvaneHimia7.htm), these standards usually refer to the levels "average, good, very good, excellent", but a performance standard for the level "fail 2" is missing. Usually it is implicitly presumed that the student is at level "fail 2" just when he is not at any of the higher ones.
Unfortunately, these limitations of the existing system for assessment in Bulgaria would hamper, not only the application of the methods that are the subject of this research but also of all scientific methods for setting cut scores. The only possible solution of this problem is that, when in each concrete case it is necessary to set cut scores for transforming the test (examination) score into a six-point marking scheme, this is implemented after first setting the corresponding performance standards.
To sum up, if we go back to our illustrative example, since the expected level of language competence of the tested seventh grade students is at A2/B1/B2, the purpose will be to set **two** cut scores. In this way, depending on their test score, the examinees will be divided in **three** classification groups (levels of language competence): **≤ A2** | **B1** | **≥ B2**.

2.2.2 Judgment

Judgment is the main and immutable part of each method for setting cut scores. Because of a half-century of experience, certain requirements have been proposed concerning the number of participants who have to take part in the judgment process, their selection and their preliminary training (Reid, 1991; Fehrmann et al, 1991; Berk, 1995; Berk, 1996; Norcini, & Shea, 1997; Hurtz & Hertz, 1999; Raymond & Reid, 2001; Clauser et al, 2002; Ferdous & Plake, 2005; Hambleton & Pitoniak, 2006; Cizek & Bunch, 2007 and others). These three factors (number, selection and training of the judges) play a significant role in setting cut scores. That is why, in the comparative analysis of several methods, it is of great importance to neutralize or to control their influence. The easiest way for controlling the influence of these factors in setting cut off scores is using in the judgment process the same judges who have gone through the same course of training.

2.2.2.1 Assigning the judgment task

The formulation of the judgment task itself is the other factor on which the final cut scores depend. In this particular case, however, this factor is not significant since all methods are based on the same judgment task.
In other words, what distinguishes the different methods in the current study is the way cut scores are set on the basis of the same data derived from the judgment process.
The formulation of the judgment task itself for which the methods (the subject of the current research) is as follows:
Which is the minimal level of competence that a given examinee should have in order to answer this test item correctly?
This is the general formulation of the task which is applicable to any content domain and any system of performance standards.

In cases when it is about setting cut scores in the area of foreign language testing and the levels of language competence match those set in the CEFR, usually the following modification of the task is used: *Which is the minimal level of language competence (A1, A2, B1, B2, C1, C2), at which a certain examinee has to be in order to answer correctly this test item?*

An often used version of this formulation (eg., Council of Europe, 2003, p. 91) is: *"At what CEFR level can a test taker already answer the following item correctly?"*

This formulation of the judgment task is used both on test items with a dichotomous scoring (correct/incorrect) of the answer and on test items with polytomous scoring or different weights. In case of polytomous scoring, the judgment is given for the maximum scoring category of the test item (completely correct answer). In cases where the test items have different weights in determining the final test score, this weight is taken into account only in setting the final test scores, but not in the judgment process.

On the basis of the thus formulated judgment instruction, the judge has to state for each item the minimum level of competence at which a given examinee has to be in order to answer the item correctly. In other words, the judge has to answer the question formulated in the judgment task as many times as there are items in the test. As a result, each judge classifies the test items into several groups the number of which is smaller than or equal to the levels of competence given in advance.

This formulation of the task matches to a large extent the one used in the Livingston (1991) method, which also requires classification of the items by the level of competence needed for answering them correctly.

Other methods in which the final result from the judgment process is a similar classification of the items by levels is the original formulation of Angoff's method (Angoff, 1971, p. 514) and its modifications (Loomis & Bourque, 2001; Impara & Plake, 1997), Jaeger's method (Jaeger, 1982, p. 463) and the item-description matching method (Cizek & Bunch, 2007, pp. 193-205).

There are two main advantages of this formulation:

First, it does not require evaluation of the probability of a correct answer for the items. The evaluation of the probability for a correct answer is considered a main weakness of the most widespread test-referenced method for setting cut scores – Angoff's modified method (Angoff, 1971, p. 514), since this is a *"nearly impossible cognitive task"* for the judges (Pellegrino et al, 1999, p. 166) and they do not always manage to accomplish it well enough (Brandon, 2004).

Second, in most test-centered methods, the judgment task requires an evaluation of a given item for examinees who are located exactly on the borderline between two levels of competence. The conceptualization of this borderline competence (at the border between two levels) usually taxes the judges even more (Berk, 1986; Norcini, 1996) than the conceptualization of a given examinee who is at a given concrete level of competence such as is required by the formulation of the judgment task for all the methods in the current study. Of course, this difference in the target group (at the borderline between two levels and at a given concrete level) leads to significant differences both in the interpretation of the results from the judgment task and in setting the respective cut scores, but these differences will be explored later on.

2.2.2.2 Judges

The most frequent answer to the question "What is the minimum number of judges that have to take part in setting cut scores for a given test?" is: the more the better, because this will lead to minimizing the standard error. There are, however, some more concrete requirements about the minimum number of judges which have emerged from experience and concrete scientific research. For instance, Livingston and Zieky recommend the number of judges to be no less than five (Livingston & Zieky, 1982, p. 16). Based on cases from forensic practice and from his own experience, Biddle (1993) suggests that 7-10 judges are sufficient. Based on the analysis using *generalizability theory*, Hutz and Hertz (1999) recommend 10 to15 judges. The recommendation given by the pilot manual for relating examinations to the CEFR is similar: at least 10 judges (Council of Europe, 2003, p. 94). However, these recommendations are related to real situations when decisions influencing the future of the examinees will be made on the basis of the set cut scores. Since the present case concerns only an illustrative example, which aims to provide a description of the different methods for setting cut scores, only one judge is needed. One judge will be enough because these cut scores will not be used in making any decisions concerning particular individuals.

Column 9 (E1) of the table in Appendix 2 represents the results from the judgment by this "live" judge. In representing the data, the following coding is used: 1 – A1; 2 – A2; 3 – B1; 4 – B2; 5 – C1; 6 – C2. In other words, if the number 3 is assigned to a given item in column E1, this would mean that, according to the judge, this item can be answered correctly by examinees whose level of language competence is B1.

The coding using numbers is applied in this case not only to facilitate the representation of the data, but also because the result from the judgment task leads to **ordinal scaling** of the test items. This happens because it is assumed that the levels of language competence form an ordinal scale. In other words, if two examinees have been rated to be at different levels of language competence, the one who has the higher level is more competent than the other, i.e. A1<A2<B1<B2<C1<C2. This assumption concerns not only levels of competence defined by the CEFR but also all performance standards used in achievement testing. However, when necessary, ordinal scaling can be dichotomized at each distinct level of competence. Such dichotomization is necessary both in some of the methods that are the subject of the study and in the analysis of their main characteristics. That is why Appendix 2 also includes the results from the dichotomization at levels A2, B1 and B2. For the rest of the levels, the dichotomization is not presented because is not necessary for the present study and, besides, this can always be done on the basis of given judgments.

The dichotomization itself is done by assigning '1' (yes) to all items which, according to the judge, can be answered correctly only by examinees at a given or lower level, and '0' (no) to all items that can be answered correctly only by examinees which are at a higher than the given level. For example, the minimum level of language competence at which a given examinee has to be in order to provide a correct answer to the first item is B2 (4). In other words, according to the judge, this item cannot be answered correctly by examinees who are at levels lower than B2 (A1, A2 and B1). Hence, in dichotomizing this item for levels A1, A2, B1, '0' has to be assigned whereas it is coded as '1' in dichotomizing for level B2 and the higher levels (C1 and C2).

An important logical conclusion, which can be drawn on the basis of the dichotomization, is that ranking the items as a result from the judgment of a given judge is, in fact, ranking the items by their difficulty according to the subjective perception of the item difficulty. This is because each item, classified by the judge at a lower level of competence, should be answered correctly by all examinees at the higher levels of competence as well. Hence, the lower the level at which the given judge would assign the item, the more examinees will be able to answer it correctly (i.e. the percentage of correct answers will be higher) and hence it will have lower difficulty, at least in the estimation of that judge.
This important logical conclusion leads to establishing one of the most important criteria for the validity of the judgment, namely the degree of agreement between the judgment and the empirical data (*intra-judge consistency*).
To illustrate the influence of this factor (intra-judge consistency) on the final value in the different methods, in addition to the judgment of the real judge E1 in Appendix 2, the judgments of one more (fictitious) judge – E2 – are provided. The intra-judge consistency between judge E2 and the empirical data is set at a maximum. In the case of this fictitious judge (E2), the same frequency distribution of items by levels is maintained as with the real judge (level A2 – 9 items, level B1 – 12 items, level B2 – 4 items and level C1 – 2 items), but the items are distributed on the respective levels depending on their difficulty (*Z*), as the eight items with a lower difficulty are assigned to level A2, the next 13 in terms of their difficulty at B1, and so on. The ranking of this fictitious judge and the corresponding dichotomizations are presented in Appendix 2 – columns 10, 12, 14 and 16.

2.2.2.3 Intra-judge consistency

The term *'intra-judge consistency'* (consistency between the judgments and the empirical data) was introduced by van der Linden (1982) and although it has other meanings as well[6], in the current study it will be used only for designating the degree of agreement between the judgments and the empirical data. The reason for this is that there is only a single judgment (conducted only once) in all the methods that are the subject of this research.
The relatively low degree of consistency between the judgments and the empirical difficulty of the items is one of the main disadvantages in most test-centered methods for setting cut scores. The accumulated empirical data over the years shows that judges usually meet difficulties in predicting the behaviour of the test items for the different groups (different levels of competence) of examinees (Reid, 1991; DeMauro & Powers, 1993; Impara & Plake, 1998; Goodwin, 1999; Pellegrino et al, 1999 and others). The factors that influence the degree of consistency are many and varied. The most significant among them are the judges' prior experience in the area of testing and test item construction as well as the duration and the nature of the training activities before the actual judgment session (Norcini, J. et al., 1988; Plake et al, 1991; Reid, 1991; Chang et al., 1996; Ferdous et al, 2006 and others).
On the other hand, the consistency of judgments with the empirical data is one of the main criteria for internal validity and the adequacy of the cut scores because this is one of the

6 For example, it can be used to refer to the degree of consistency of the judgments from a single judge in different rounds of setting cut scores.

few possibilities for "objectifying the subjectivism" through its linking with reality (Kane, 1994; Messick, 1994; Messick, 1995; Linn, 1998; Hambleton & Pitoniak, 2006; Reckase, 2006; Cizek & Bunch, 2007).

Figure 10 is a visual illustration of how consistent the judgment by the first judge (E1) is with the test item difficulty obtained from the empirical data. On this graph, each test item is represented as a point with coordinates: ***x*** – is the empirical difficulty, expressed on the z-scale[7] and ***y*** – the minimum level of competence at which the item will be answered correctly, according to the judge (E1).

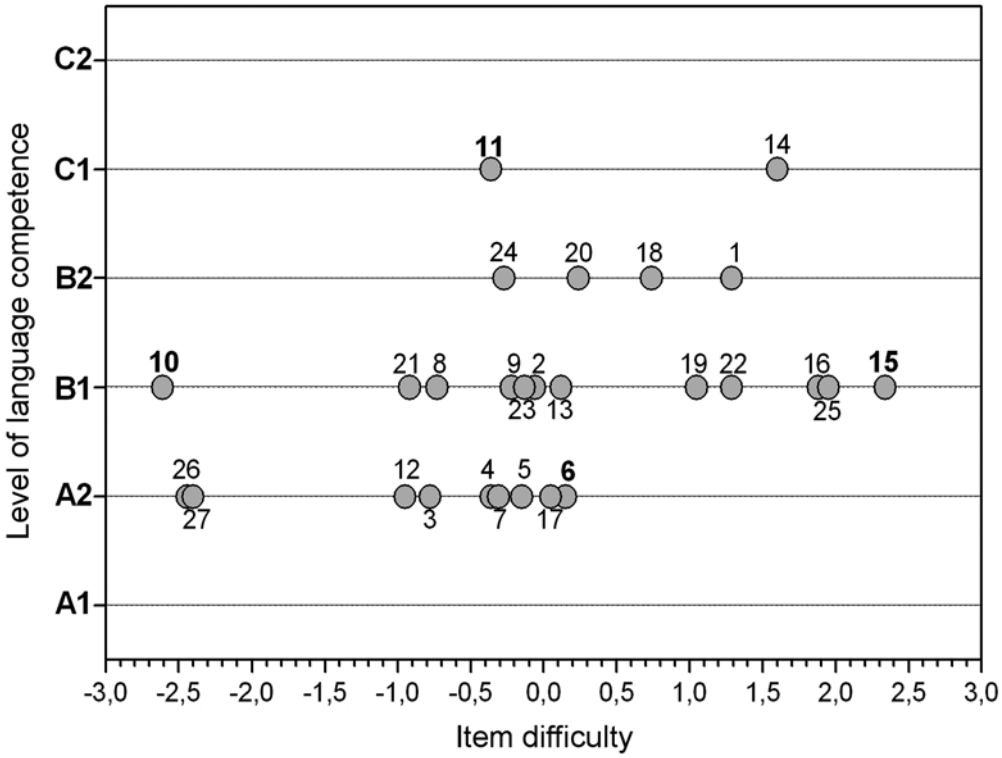


Figure 10 Consistency between the judgment and the empirical data – real case (E1)

Figure 10 shows that the most difficult item, № 15 (z_{15} = +2.34), which was answered correctly by only 16% of the examinees, should be able to be answered correctly by an examinee who is at level B1, according to the judgment of the real judge (E1). If this item can be answered correctly by someone who is at level B1, then anyone who is at the higher levels (B2, C1 and C2) should also answer it correctly. Hence, the percentage of the correct

7 Item difficulty is represented on the z-scale since it is the same for the two types of data (real and simulated), as distinct from the percentage of the correct answers whose values, although close to each other, are different for the two samples.

answers for this item should be higher than the percentage of correct answers for item № 11 which, according to the judge, can be answered correctly only by examinees which are at level C1 (and C2). In reality, however, item № 11 is answered correctly by 56% of the students, and this is more than three times the percentage of correct answers to item № 15.

On the other hand, the item with lowest difficulty in the test, № 10 (z_{10} = -2.61), can be answered correctly, according to the judge, at level B1 (and all higher levels) and therefore should be more difficult than item № 6 (z6 = +0.15), which can be answered correctly, in the judge's opinion, not only by those who are at level B1 (and all higher levels), but also by the examinees who are at level A2. The empirical data, however, shows that the judge is wrong again: the percentage of t correct answers for item № 10 is 88%, and for item № 6 it is 48%.

These inconsistencies between the judgment and the empirical data are not the only ones, as can be seen in Figure 10, but they show clearly how serious the problem is. In the ideal case, the distribution of the items by level of competence has to be such that the items with lower difficulty have to be assigned to the lower levels and the ones with higher difficulty to the higher levels of competence, as in the case with the fictitious judge E2 (Figure 11). Figure 11 shows clearly that the distribution by levels of competence is in perfect match with their difficulty. Unfortunately, in real life no perfect things do exist. Moreover, it has to be underlined that the real judge (E1) is one of the leading judges in foreign language testing in Europe in linking the tests with the CEFR, with experience of many years both in the construction of test items and in foreign language education.

As will be seen later, in real situations the degree of consistency between the judgments and the empirical data for the majority of judges is even lower than the one presented in Figure 10.

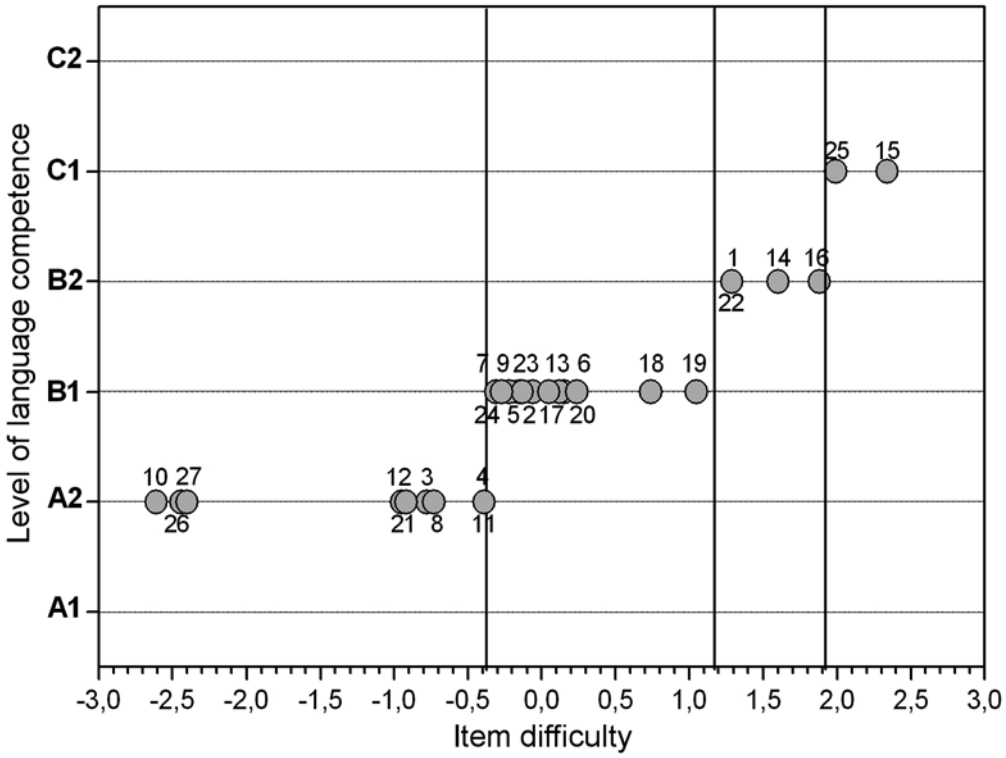


Figure 11 Consistency between the judgment and the empirical data – the ideal case (E2)

Since a complete match between the judgments and the empirical data is hardly achievable in reality, the aim is at least that the mean difficulty of the items, assigned by a given judge to the same level, is an increasing function of these levels. In other words, the mean difficulty of items, classified by the judge as belonging to a lower level, should be lower than the mean difficulty of the items assigned to the higher levels. This condition is satisfied for both judges in this illustrative example, as can be seen in Figure 12. On the basis of the graphical analysis (Figure 12a and Figure 12b), several main conclusions can be drawn.

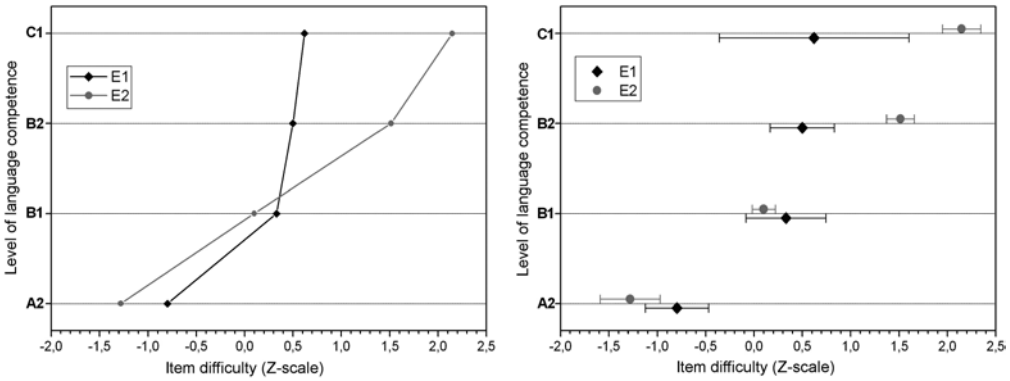


Figure 12 Judgment: mean difficulty of the items by levels of competence (a) and standard errors of the means (b)

First, the mean item difficulty is a strong growing function of the levels of competence in both judgments. In other words, both judges fulfil the main criterion for the degree of consistency between the judgments and the empirical data.
Second, there is a big difference in the mean difficulty of items between the two judgments and the tendency is different for the different levels. For the lower levels (A2 and B1), the mean difficulty of items assigned by E1 to these levels is higher than the mean difficulty of those assigned by E2, but the difference in the mean values is statistically not significant. For the higher levels (B2 and C1), however, the mean item difficulty of the items is higher for E2, and the difference in the mean values between the two judges is statistically significant. These differences will inevitably reflect the cut scores later, if they are based on the judgments of the two judges, and this is expected at least in most of the analysed methods.
Third, the differences between the mean values for the levels are smaller for the real judge (E1), and the differences between levels B1 and B2 (0.17) and levels B2 and C1 (0.12) are smaller than even the minimum standard error of the measurement (0.48). In addition, for the first judge, the differences between the mean values for the item difficulty for levels B1/B2 and levels B2/C1 are so small that they are not statistically significant. The reason for this, to a large extent, is the small number of items at these levels (four for B2 and two for C1), but this minimal number did not influence the second (fictitious) judge, in which case the differences in the mean item difficulty for each pair of adjacent levels are statistically significant.
The graphical analysis of the data is neither the only one possible nor the best method for exploring intra-judge consistency. Another appropriate and frequently applied method is correlation analysis. Pearson's coefficient of correlation is the most frequently used quantitative index for the consistency between the judgment and the empirical results. Using this coefficient is justifiable only when both variables (item difficulty and judgment in this case) are normally distributed interval-scale variables. In the formulated judgment

task the results of the judgments are expressed in an ordinal and not in an interval scale, and thus Pearson's coefficient is not appropriate.

The ordinal scale in which the results are expressed suggests that, instead of the Pearson correlation coefficient, some of the coefficients of rank correlation can be used. The two most frequently used coefficients of rank correlation are Spearman's (ρ) and Kendall's tau (τ). Spearman's coefficient, which is analogous to Pearson's coefficient for ordinal variables, is relatively widespread although, in contrast to Kendall's tau (τ), it does not have a clear and simple interpretation.

Kendall's coefficient can be interpreted as the difference between two proportions: in this case the proportion of the test items which are ranked in the same way in both variables (empirical and judgmental item difficulty) and the proportion of items ranked in a different way in both (Cliff, 1996, p. 333). Kendall's coefficient is more appropriate than Spearman's especially in relatively small samples, as is usual with samples of test items. The reasons for this are its specific characteristics in small samples – normal distribution of the statistics with no misplaced evaluation of the respective parameter, which distinguishes it from Spearman's coefficient (Lapata, 2006, p. 474).

In using these two coefficients, however, it has to be taken into account that although both vary within a limited interval [-1; +1], they do not reach maximum values in cases when the number of different ranks in the two variables is not the same. Such is the case in comparing the ranking of the empirical difficulty and the judgments. In the judgments, the number of the different ranks is limited by the number of the predefined levels of competence and this number is usually much lower than the number of the test items. Besides this, it has to be taken into account that although the two coefficients vary in the closed interval [-1; +1], their values are not comparable because they are expressed in different scales, and usually for the same samples Spearman's coefficient (ρ) is higher than Kendal's (τ). The ratio ρ/τ is approximately 3/2. (Fredricks & Nelsen, 2007).

One additional limitation in using these coefficients is that there are no criteria for their minimum permissible values for an acceptable degree of consistency between the empirical data and the judgments. The statistical significance of the respective correlation coefficient can be taken as a tentative criterion. The logic of this suggestion is that if the respective coefficient is not statistically significant ($\neq 0$), then there is no association between the two variables.

As the data in Table 4 show, for the real judge (E1) both rank order correlations with item difficulty, although relatively low, are statistically significant ($\rho = 0.43$; $\tau = 0.32$) with a probability for error lower than 5%. In other words, the intra-judge consistency between the judgments of the first judge and the empirical data is acceptable since the rank correlations (ρ and τ) are statistically significant at the 5% level. The same can be argued for the fictitious judge (E2) as well, but this is logically grounded by the way the judgment was obtained.

Table 4 Indices for intra-judge consistency[8]

Index	Item difficulty / Judgment (E1)						Item difficulty / Judgment (E2)					
	Z		**% (*n* = 250)**		**% (*n* = 5000)**		**Z**		**% (*n* = 250)**		**% (*n* = 5000)**	
ρ[1]	**+0.43**	(.025)[2]	**-0.43**	(.025)	**-0.43**	(.025)	**+0,93**	(.000)	**-0,93**	(.000)	**-0,93**	(.000)
τ_b[3]	**+0.32**	(.035)	**-0.32**	(.035)	**-0.32**	(.035)	**+0,83**	(.000)	**-0,83**	(.000)	**-0,83**	(.000)
A2: r_{pb}[4]	**-0.45**	(.018)	**+0.46**	(.016)	**+0.47**	(.014)	**-0,72**	(.000)	**+0,74**	(.000)	**+0,74**	(.000)
B1: r_{pb}	-0.23	(.248)	+0.24	(.222)	+0.24	(.220)	**-0,73**	(.000)	**+0,74**	(.000)	**+0,75**	(.000)
B2: r_{pb}	-0.14	(.486)	+0.14	(.482)	+0.14	(.500)	**-0,49**	(.010)	**+0,47**	(.013)	**+0,47**	(.013)
MPI[5]	**0.698**		**0.698**		**0.694**		**+1.00**		**+1.00**		**+1.00**	

1 Spearman's coefficient
2 Confidence probability (p), the zero before the decimal sign is omitted to save space. At $p < 0.05$ the corresponding correlation coefficient is statistically significantly different from 0.
3 Kendall's coefficient
4 Point-biserial coefficient
5 Index of intra-judge consistency

The data in Table 4 also confirm that the values for Spearman's coefficient are higher than the ones for Kendall's coefficient. Moreover, although the judgment for the second judge (E2) was formed in a way to reflect a perfect judge, the two coefficients of rank correlation ($\rho = 0.93$; $\tau = 0.83$), although considerably higher than those for E1, do not reach the maximum value of +1 due to the different number of ranks in both variables.
Concerning the rank correlation between the judgments and the item difficulty expressed as a percentage of correct answers, the results in Table 4 show clearly that this correlation is almost the same as with the difficulty parameters. The substantial difference is only the sign for the correlation, which is negative for the difficulty (expressed in percentages), because a high percentage correct answers corresponds to an easy item and vice versa.
The dichotomization of the judgments by levels allows, in addition to using the coefficient of rank correlation described above, the use of the point-biserial correlation coefficient (r_{pb}) as an index of the degree of consistency between the judgments and the empirical difficulty of the items. This correlation coefficient is a special case of Pearson's coefficient of correlation when one of the variables is dichotomous. As it can be seen in Table 4, the point-biserial coefficient of correlation does not reach its maximum even for the perfect judge. The main limitation of this coefficient, however, is that in its application not just a single but several indices of consistency are derived (with one less than the number of the different levels of competence used in the judgment task). In this case, some indices can be statistically significant and others not, as in the case for the real judge (E1). For level A2 the

8 Editors' note: the meaning of the column heads in the table is a follows. 'Z' reports the correlation between judgment and difficulty parameters; '%' reports the correlations between judgment and percentage correct.

point-biserial coefficient of correlation for judge E1 (r_{pb} = -0.45[9]) is statistically significant at 5% level of confidence while for the next two levels (B1: r_{pb} = -0.23; B2: r_{pb} = -0.14) this is not the case. This means that if only this index for determining the consistency between the judgments and item difficulty is used, it will not be possible to draw an unequivocal conclusion about the adequacy of the judgments.

If, however, the point-biserial coefficient of correlation is used during the training of the judges, the multiple measures of consistency can turn into an advantage. The reason is that it becomes possible to determine at which levels of competence a given judge meets difficulties. Consequently, the training can focus mainly on the proper conceptualization of these levels.

Although rarely applied in real situations for setting cut scores, there are other methods for determining intra-judge consistency as well. (Van der Linden, 1982; Bay & Nering, 1998; Wiliam, 1998 and others). All these methods, however, have the same limitations (more than one index for a given judge) and this in turn means that the conclusion about the degree of consistency for a given judge will not always be unequivocal.

In order to meet the needs of the standard setting methods described in this study, a new index of intra-judge consistency was developed in 2006 by the author – the *misplacement index* (MPI). This index was applied in several real situations of setting cut scores, but this is the first time its use has been reported.[10]

The misplacement index (MPI) presupposes ordinal scaling for the two variables with or without tied ranks. In its logic it is close to Kendall's rank correlation coefficient, but as it will be demonstrated, it can reach its maximum value of +1 if a different number of ranks is used in both variables. However, the main advantage of the MPI is that it allows obtaining an individual index of consistency both for given judge as a whole and for each separate test item. This, in turn, allows an in-depth analysis of the items and of the item-related factors that influence the degree of consistency.

9 The sign is negative because in the judgments a lower value (0) is assigned to the items which cannot be answered correctly at the given level, i.e. they are more difficult. In the Z-scale, however, the higher values correspond to the more difficult items.

10 Editors' note: Since the time of writing the MPI index has been reported by Kaftandjieva (2009a), Kaftandjieva (2009b) and Moe (2008).

The formula, by which the index of consistency for a given judge is calculated, is as follows:

$$MPI = 1 - \frac{\sum_{i=1}^{N} W_i}{\sum_{j=1}^{k} n_j \cdot (N - n_j)}$$

where:
N – total number of items;
k – number of levels of competence;
n_j – number of items at level j;
W_i – number of discrepancies for item[11].

In contrast to the correlation coefficients, described above, this index of consistency has values in the interval [0; 1]. It reaches its maximum when the ranking of the items is in full correspondence with the ranking of the items by their difficulty, as in the case of the second "perfect" judge (E2) for whom the index of consistency will reach the value of one. A minimum value of 0 will be obtained if the judge ranks the items by level of competence in a reverse order, putting at the highest levels the items with the lowest difficulty and the most difficult items to the lowest levels, respectively.
The index of consistency cannot be calculated in cases when, according to a given rater, all test items belong to the same level. This situation is quite possible in reality, but in such cases, no cut score can be set. If it is not possible to set a cut score, the issue of consistency between the judgment and empirical data becomes irrelevant.
The logic of calculating the index of consistency can be illustrated through a specific example. Let us take item № 1, whose difficulty is +1.29 (with a percent-correct rate of 29% and 28%, respectively). If we rank the test items according to their difficulty, putting the one with the lowest difficulty (№ 10) in the first position and the item with the highest difficulty (№ 15) in the last position (27), then item № 1 will be placed in the 23rd position (i.e. 22 of the test items are easier and only 4 items are more difficult than it). Figure 13 shows this ranking of items by their difficulty. Each item is represented by a circle featuring the level of competence at which each judge (E1 and E2) had placed this item. The levels are designated by numbers as follows: 2 – A2, 3 – B1, 4 – B2, 5 – C1.
According the real judge (E1), the minimum level of language competence at which a given examinee has to be in order to answer correctly item № 1 is level B2 (= 4). In Figure 13, in the row corresponding to the judgment of the first judge (E1), this item (№ 1) is marked in grey and its position corresponds with its difficulty. The number of the item is shown on the x-axis.

11 Editors' note: If item 1 is objectively easier than item 2, but is judged to be more difficult, this is a discrepancy. This discrepancy has to be counted as a discrepancy for item 1 and also for item 2.

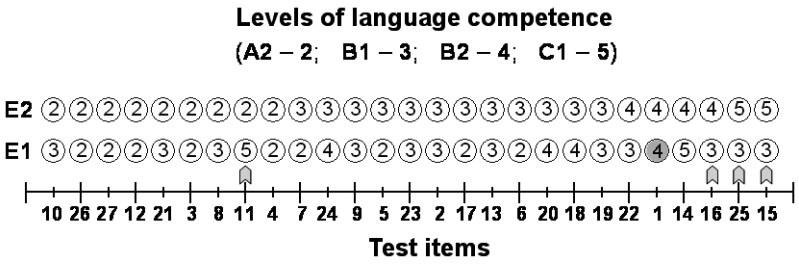


Figure 13 Index of correspondence (MPI) – item ranking

If we examine the 22 items which are located to the left of item № 1, we will see that the level of rated competence for one of them (item № 11), which is marked with an arrow, is higher than the level of competence for item № 1 (5 > 4). This is one of the disparities that are taken into account in calculating the index of consistency. The other three disparities are for the last three items in the ranking by judge E1 (items № 16, № 25 and № 15). Although their empirical difficulty is higher than in № 1, according to the "real" judge they can be answered correctly at a lower level of competence (B1 = 3) than item № 1 (B2 = 4). In other words, the total number of discrepancies for item № 1 is four (W_1 = 4). The number of the items that this judge has placed at level B2 is also four.

Repeating this procedure for all test items and summing the corresponding numbers of discrepancies for each item we will find the number that is equal to the numerator for the right side of the formula for calculating MPI. In this case, for E1, the numerator is equal to 146.

The denominator is equal to the total number of possible discrepancies that would appear if the judge ranked the items in a descending order by their difficulty. With 27 items and four different levels of competence used in the judgment, for the "real" judge this number of possible discrepancies is equal to 484.

After substituting these values in the formula and doing the necessary calculations the final result is obtained, namely: MPI_{E1}= 0.698. The interpretation of this result is unequivocal – the value of the index of consistency (MPI) is equal to the relative share of the correct ranks, i.e. in nearly 70% of the cases the ranking of the items by the first judge corresponds to their rank order as defined by the empirical difficulty of the items.

This interpretation of the index of consistency logically leads to the **minimum acceptable criterion** for the degree of consistency between the judgments and the empirical difficulty of the items. According to this criterion, the **judgment is acceptable, if MPI > 0.5**, i.e. when in more than half of the cases a correct ranking of the items is achieved. In other words, different from the correlation coefficients, where the critical value is defined on the basis of statistical significance, in MPI the minimum criterion is based only on its interpretation and the principle of majority.

Using this criterion, the judgment of the real judge can be estimated to be acceptable in regard to its consistency with the empirical data.

However, in setting cut scores and for decisions concerning individuals, it is preferable – instead of the minimum criterion – to set a **more exacting criterion (MPI > 0.70)**. This

criterion is congruent with the critical values (Karabatsos, 2003; Harnish & Linn, 1981) of *modified Sato's caution index* and is close to the criterion for consistency between two judges in judging open-ended items (> 70% absolute agreement).

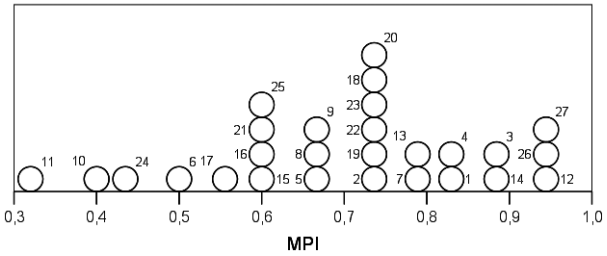


Figure 14 Frequency distribution of the items by index of consistency (MPI_{EI})

As was already mentioned, one of the main advantages of this index is that it can be calculated separately for each item. If, for example, we take again item № 1, we noted that the number of the discrepancies for it is equal to four. Since this item was placed by the judge (E1) at level B2 where he placed three more items (n_k = 4), the maximum number of discrepancies for this item is 23 ($N - n_k$ = 27 – 4 = 23). Therefore $MPI_{E1\text{-}1}$ = 1 – 4/23 = 0.83, which is a relatively good index of correspondence, especially compared to item № 11, which has the lowest index of correspondence (0.32). As can be seen in Figure 14, where the frequency distribution of the items is presented in terms of their index of consistency, there are three items (№ 11, № 10 and № 24) for which the number of the incorrect rankings is higher than the number for the correct rankings.

The comparison of the average index of consistency of the items for the real judge confirms, to a certain extent, the finding from the analysis of the point-biserial coefficient of correlation. It can be shown that the most adequate are the judgments for the items which this judge places at level A2. Additional support for this finding derives from the average index of consistency for this level which exceeds the average indices for the rest of the levels ($MPI_{E1\text{-}A2}$ = 0.78; $MPI_{E1\text{-}B1}$ = 0.66; $MPI_{E1\text{-}B2}$ = 0.68; $MPI_{E1\text{-}C1}$ = 0.60), although the difference between the mean values is not statistically significant, mainly because of the small sample sizes. As a whole, however, the mean value of MPI is higher than 0.5 for all levels, i.e. the judgments of E1 can be accepted as adequate for all levels of competence indicated by him.

In the case of the second judgment (E2), since it was obtained with a perfect consistency with the items' empirical difficulty, the index of correspondence reaches its maximum: MPI_{E2} = 1 (Table 4)[12].

12 Editors' note: We wish to point out that there is a tacit assumption that there are no tied ranks. While this may normally be the case tied ranks are possible and then the formula is not fully adequate.

2.3 Setting cut scores

In this section several different, relatively new methods for setting cut scores will be described. These methods use the same available data. Nevertheless, the processing of these data is different and leads to different cut scores.
The methods themselves were developed in the period between 1999 and 2006 for the needs of foreign language testing, and especially for linking the test results to the levels of language competence as defined in the CEFR. Although used many times in real test situations, these methods have not been examined in detail and there is still a lack of scientific publications providing arguments supporting their validity. That is why the present study, in addition to describing the methods, also reports the results of an analysis of their major characteristics and their advantages and disadvantages.

2.3.1 Basket procedure

This method was created and applied for the first time during the second phase of the international European project for Internet-based foreign language testing – DIALANG (http://www.dialang.org/english/index.htm) in 2001. Subsequently it was also used in several Dutch projects for setting cut scores (eg. Noijons & Kuijper, 2006). A brief (and to some extent inaccurate) description of the method is given in the pilot manual for relating language examinations to the CEFR (Council of Europe, 2003, p. 91)[13].
According to this method, each cut score is equal to the total number of items which, according to a given rater, can be answered correctly at all levels of competence which are below the respective cut score.
Since the frequency distribution of the items by levels of language competence is the same for both raters (Figure 15), using the Basket method, their cut scores will be the same. (A2/B1 = 9; B1/B2 = 12; B2/C1 = 25).

	A1	A2	B1	B2	C1	C2
E1	0	9	12	4	2	0
E2	0	9	12	4	2	0

	(0)	(0+9)	(9+12)	(21+4)	(25+2)
Cut scores	**–**	**9**	**21**	**25**	**-**
	A1/A2	**A2/B1**	**B1/B2**	**B2/C1**	**C1/C2**

Figure 15 Cut scores (Basket procedure)

The logic of the method itself is simple and accessible for the general audience (the examinees themselves, their parents, teachers, etc.). According to this logic, to put a given examinee at X-level of competence, this examinee has to answer correctly not only the items belonging to the lower levels of competence but at least one more item which is at

13 Editors' note: A substantially revised version of the Manual is available at: http://www.coe.int.

this level X or at some higher level.
What should be noted here is that with this method the cut scores can be set only when the total number of items belonging to the levels preceding a certain level is different from **zero** or from the **maximum number** of test items. In other words, based on these judgments (of E1 and E2) cut scores A1/A2 and C1/C2 could not be set, even if this were indispensable.
The other important thing is that the cut scores set in this way define the upper limit of the test score for the level of competence which precedes the respective cut score, i.e. the maximum number of items a given examinee can answer correctly and still be at level of competence which is **under** this cut score. Consequently, based on the cut scores defined this way, depending on their raw cut score, the examinees can be classified by levels of competence as follows:

- Test score in the interval [0; 9] → level ≤ A2;
- Test score in the interval [10; 21] → level B1;
- Test score in the interval [22; 26] → level B2;
- Test score in the interval [27; 28] → level ≥ C1.

In this case, such a classification the levels of competence located at the two ends are left open (≤ A2 and ≥ C1) because the data does not allow inferences to be made for levels A1 and C2.
Besides this, although the two judgments of both judges allow the examinees to be classified at four different levels, bearing in mind the presumed level of language competence of the examinees (A2/B1/B2) and the pursuit for maximally precise classification, the number of cut scores will be limited to two (A2/B1 and B1/B2) which results in three classification categories, namely:

- Test score in the interval [0; 9] → level ≤A2;
- Test score in the interval [10; 21] → level B1;
- Test score in the interval [22; 27] → level ≥B2.

The transformation of the raw test score into a Z-scale in using IRT depends on the model that was applied. In applying the one-parameter Rasch model, as in this illustrative example, the transformation into a Z-scale is direct because all items have the same weight. The transformation itself is implemented by using transformation tables similar to the one in Appendix 3. According to this table, a raw test score of 9 correctly answered items corresponds to -0.85, and a raw test score of 21 correctly answered items corresponds to +1.57, and consequently the intervals for the three classification levels will be: level ≤ A2 ∈ (-4.86; -1.85]; level B1 ∈ (-1.85; +1.57]; and level ≥ B2 ∈ (+1.57; +4.73).
In other IRT probability models, instead of the raw test score, a weighted raw test score, utilizing differential item weights, has to be calculated in advance. This logic in setting the cut scores is followed also when, instead of dichotomous scoring, some items are scored polytomously regardless of whether the test is based on classical test theory or IRT.
The main **advantages** of the Basket procedure are as follows:

- Since, in contrast to most test-centered methods, there is only one round of judgments, its application requires considerably less time. Combined with its simplicity, its practicality and economic use of time make it an attractive method for many situations in which resources and technical expertise may be limited .

- Setting cut scores through the Basket procedure does not require special statistical methods or software. Thanks to this, the method is especially attractive for internal assessment and non-standardized tests developed by teachers.
- The Basket procedure has a wide range of applications because it can be used both in dichotomous and polytomous scoring of the test items. Moreover, it can be used irrespective of the approach adopted in developing the test itself (classical test theory or IRT).
- An additional quite important advantage is its clear interpretation, which allows the interpretation of the results to be made using language which is easily understandable to general audience.
- Since setting cut scores only needs the results of the judgments (the corresponding frequency distributions of the items by levels of competence), the cut scores themselves can be set in advance without conducting the test. In this sense the test can be defined as purely judgmental using Berk's classification (Berk, 1986, p. 139).

The last one of these advantages of the Basket procedure can be regarded as a disadvantage as well, depending on the point of view. The reason is that this procedure squarely contradicts one of the key recommendations in the literature on standard setting, which calls for a maximum integration of the results from the judgments and the empirical data in order to maximize the adequacy and validity of the corresponding cut scores (Livingston & Zieky, 1982; Jaeger, 1990; Norcini, 1994; Pellegrino, J. et al, 1999; Zieky, 2001; Linn, 2003 and others).

As with any other method, however, the Basket procedure also has some serious **disadvantages**, which question its validity as a whole and consequently the validity and the adequacy of the cut scores set by using it.

The main disadvantage of the method is that the cut scores set by it are not consistent with the adequacy of the judgments, i.e. its degree of correspondence with the empirical data. An illustration of this is the fact that the two sets of judgments lead to one and the same cut scores regardless of the fact that one of the judges (E2) has absolute consistency (MPI = 1) while for the other this consistency is significantly lower (MPI = 0.698). In addition, even if a given judge is completely inadequate and ranks the items in a reverse order of their difficulty, the cut scores based on his judgment will not differ at all from the rest if the item distribution by levels of competence is the same. In other words, the cut scores set on the basis of two sets of different judgments will differ from each other only if the two frequency distributions differ.

The fact that in practice the empirical difficulty of the items does not affect at all the cut scores set by the Basket procedure is rather strange because this contradicts not only the logic of the use of judgment but the logic of setting the cut scores as well. Thus, instead of solving the common problem of insufficient consistency of the judgments with the empirical data which is seen in all test-centered methods, the Basket procedure simply ignores it.

This is a rather serious shortcoming especially given that its potential popularity due to its simplicity and wide range of application could lead to a significant number of inadequate or invalid decisions being made for a number of tests.

There is a counter-argument that the results of an analysis of the consistency between the

judgments and the empirical data can be used for discovering inadequate judgments, and disregarding them in the final setting of the cut scores. However, this does not work in practice. The reason for this is that, very often in real standard setting situations, the degree of consistency between the judgments and empirical data is not analysed, precisely because the method itself does not require this. An example of this is the (apparently) only adequately documented application of the method (Noijons & Kuijper, 2006), in which the question of the degree of consistency between the judgments and the empirical data is not addressed at all.

The other serious theoretical disadvantage of the Basket procedure is that it can lead to a misplaced evaluation of the cut scores especially if these cut scores are at the ends of the interval within which the examinees' raw test scores vary. This phenomenon is known as *distortion of judgments*. According to Reckase (Reckase 2000-a; Reckase, 2006), one of the main criteria for the quality of a given method for setting cut scores is the **criterion for minimal distortion of the cut score**.

The possibility for a distortion of the cut score stems from the characteristics of the judgment task itself and its statistical interpretation. The judge's task is to define the minimum level of competence that the examinee needs to be at in order to correctly answer every single item. From a statistical point of view this means that at the appropriate level, the probability for a correct answer to the item will be higher than the probability for an incorrect answer, i.e. the probability for a correct answer (and, correspondingly, the item difficulty) will be above 50%.

Therefore, if a given judge had understood his task correctly and behaves in absolute concordance with the empirical data, he or she would assign to level Xk all the items for which the percent of correct responses is over 50% while for the preceding levels it does not exceed 50%.

Let us suppose that besides the two judges (E1 and E2) there is one more judge (E_X) who knows and accepts the cut scores set on the basis of the judgments by the other two (equal to 9, 21 and 25 points of raw test score, respectively, as the upper limit for levels A2, B1 and B2). Besides this, let us imagine the third judge also has all the empirical data. Then, if he or she understands the judgment task correctly and strictly follows the instructions, he or she will classify only three items (№ 10, № 26 and № 27) as belonging to level A2 because these are the only items for which the percent correct rate is higher than 50% for the examinees with a raw test score between 0 and 9 points (Table 5).

Table 5 Basket procedure: percent of correct responses for items at different levels of language competence

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Real data (n = 250: n_{A2} = 86; n_{B1} = 124; n_{B2} = 34; n_{C1} = 6)																											
A2	9	22	21	10	22	9	10	28	28	**76**	29	42	24	1	1	3	12	5	9	21	35	9	24	26	7	**71**	**71**
B1	31	**56**	**81**	**75**	**61**	**59**	**73**	**76**	**60**	94	**65**	**74**	**52**	23	15	17	**59**	43	33	**51**	**77**	27	**62**	**63**	18	94	94
B2	**62**	97	100	97	91	94	100	97	88	100	85	94	85	**76**	**47**	**74**	100	**94**	**79**	85	100	**76**	79	91	**52**	97	94
C1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	83	67	100	83	100	100	100	83	100	100	100	100	100
Simulated data (n = 5000: n_{A2} = 1567; n_{B1} = 2728; n_{B2} = 567; n_{C1} = 138)																											
A2	6	17	27	22	18	13	21	30	20	**70**	22	30	14	4	2	3	15	8	7	13	34	5	16	19	3	**67**	**64**
B1	28	**59**	**76**	**68**	**62**	**57**	**67**	**73**	**63**	94	**67**	**78**	**57**	24	15	21	**59**	41	33	**54**	**77**	29	**63**	**66**	18	**94**	**94**
B2	**72**	90	95	94	93	90	92	97	94	100	93	97	91	**69**	**53**	**60**	90	**84**	**78**	89	97	**73**	93	93	**58**	100	99
C1	94	99	100	99	100	99	100	100	99	100	100	100	96	98	85	91	98	99	99	99	99	96	100	99	87	100	100
(n = 5000: n_9 = 220; n_{10} = 228; n_{21} = 206; n_{22} = 157; n_{25} = 112; n_{26} = 89)																											
9	9	29	49	40	34	23	36	45	32	83	39	51	24	8	4	7	32	13	11	23	54	12	29	35	8	85	80
10	12	32	52	42	34	30	43	51	41	89	46	61	31	11	4	7	32	21	13	30	59	12	36	44	8	88	85
21	52	80	93	90	88	83	86	93	88	99	83	92	81	47	36	46	84	69	62	82	91	57	88	86	39	99	98
22	63	85	93	91	91	86	90	95	91	100	90	93	88	64	36	46	86	80	70	87	96	57	91	90	44	100	98
25	86	96	97	96	96	96	97	99	95	100	96	100	94	79	65	79	95	89	89	95	99	84	99	96	77	100	100
26	91	98	100	99	100	98	100	100	99	100	100	100	94	97	76	87	97	98	99	98	99	93	100	99	80	100	100

Following the same logic, the hypothetical judge (E_X) should assign at level B1 16 items (№ 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 17, 20, 21, 23, 24), and at level B2 8 items (№ 1, 14, 15, 16, 18, 19, 22 and 25). At level C1 the judge E_X should not assign any item because the minimum level at which every single item is mastered is lower than C1[14]. The cut scores set on the basis of judgments by E_X using the Basket procedure method would be 3 and 19, respectively. These values are definitely different from those on whose basis the examinees were grouped by levels of competence and which the hypothetical judge tried to reproduce by following the instructions and taking into account all the empirical data. Moreover, there is one less cut score: the judgments by E_X do not allow setting a cut score between levels B2 and C1 because all items of the test can already be answered correctly by the examinees at B2.

It should be noted that the cut scores set on the basis of the judgment by E_X will be the same regardless of whether the judge is using the real or the simulated data. The simulated data, however, due to the bigger sample size, allow defining the item difficulty not only for each particular level, but also for each particular test score. For example, in the lower part of Table 5, the percentage of correct item difficulties for each item for examinees with test

14 Editors' note: The author has made a mistake here. In the upper part of Table 5, the difficulty of item 15 at level B2 is only 47 (% correct) and therefore judge EX should not have put this item in basket 'B2' but in basket 'C1'. This mistake, however, does not invalidate the author's general reasoning. For the simulated data (middle part) the reasoning is correct. The point made applies, however, also to the claim made in the first sentence of the next paragraph.

scores 9, 10, 21, 22, 25 and 26 are provided. These values are selected because they match the exact borders (right and left) of the separate levels of language competence as set on the basis of the judgments by E1 and E2. As can be seen in Table 5, for none of the raw test score set in this way, does the number of items with a difficulty of 50% or higher match the number of item answered correctly by these examinees.

This phenomenon is called **distortion of judgments**. It is mostly evident in the alternative Angoff method and its modifications. (Reckase, 1998; Hambleton & Pitoniak, 2006; Cizek & Bunch, 2007). The main reason for this distortion of the cut scores is in the formulation of the judgment task which, instead of evaluating the probability (from 0 to 100 %) for each correct answer, requires a dichotomous response in terms of "**is able/is not able** to answer given item correctly" at the respective level of competence. This simplification of the judgment task leads to a subsequent misplacement of the respective cut scores (Reckase, 1998, p. 10).

Since the judgment task in the Basket method is quite close in its formulation to the judgment task in the alternative Angoff method, the misplacement of the cut scores is obvious in it as well.

Figure 16 shows the misplacement of the cut score for each possible value of the test score when the Rasch model is applied.

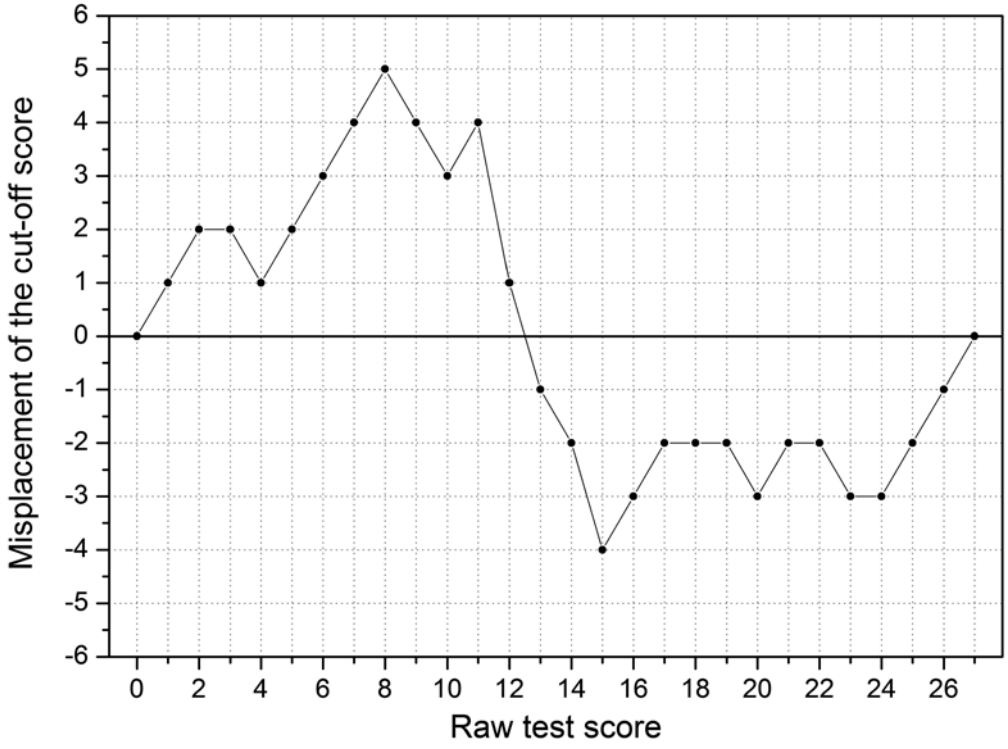


Figure 16 Misplacement of the cut score (expected and obtained) in the Basket procedure

The line on this figure shows the difference between the expected (the test score) and the obtained cut score when applying the Basket procedure for the data from hypothetical judgment data which are fully congruent with the empirical data.
In an ideal case, all the differences should be equal to 0, i.e. all points on the figure should be on the thick horizontal line in the middle. Unfortunately, except the two extreme values (0 and 27), considerably high misplacement of the cut score is apparent (on average more than two points: 2.29).
Moreover, except for the two extreme values (0 and 27), there are no cases of agreement between the expected and the actual cut score set on the basis of the hypothetical, ideal judgment data.
There is also a clear tendency in this misplacement. According this tendency, with a test score under the average for the sample under consideration, the cut scores are lower than expected, and with test-score over the average the cut scores are higher than expected.
If the misplacement of the cut score is expressed as a percentage of the maximum possible difference (27), then the mean misplacement in absolute value (2.29) will be equal to 8.5%. Since this misplacement is higher than the standard error of measurement (SEM = 2.00), then this method does not comply with the criterion for the minimum distortion of judgments.
The main conclusion that can be drawn at the end of this description of the method is that although very attractive from a practical point of view, it has a lot of shortcomings which call into question its validity and the adequacy of the cut scores obtained by using it. In other words, simplicity is at the expense of its quality.

2.3.2 Compound Cumulative method

The Compound Cumulative method was created and applied for the first time in 2001 for setting cut scores and linking the results of the matriculation examinations in English in Finland to the CEFR (Kaftandjieva & Takala, 2002). Subsequently it was used many times in a set of projects in foreign language testing in several European countries (Finland, Norway, and Spain).
The goal in the development of this method was that it should correspond to three main criteria: practicality, wide range of applicability and maximum agreement with the empirical data.
Practicality in this case refers to a judgment with a minimum cognitive complexity and requires a minimum expenditure of time. This criterion to a large extent determined the format of the judgment as well as the decisions for its rounds (one) with only one purpose – to save time.
The wide range of **applicability** meant that the method should be suitable for:

- all kinds of tests, regardless of their theoretical basis (classical test theory or IRT);
- all kinds of test items, regardless of their scoring (dichotomous or polytomous).

As mentioned before, the main problem in most test-centered methods is the low level of **agreement between the judgments and the empirical data**. That is why in developing the method, an approach was sought which would minimize as much as possible the effect of this inconsistency on the cut scores. Solving this problem was also hampered by the fact that the judgment task was conducted once and this excludes the possibility to provide the

judges with empirical data and hence with the opportunity that they could/would revise their initial decisions. The only possible way to solve the problem in this particular case was aggregating the results of the judgment and the empirical data a posterior and setting the final cut scores as a result of this aggregation.
Figure 17 illustrates this a posterior process of aggregating the results from the judgments and empirical data. The test items are represented by grey circles whose horizontal position corresponds to their empirical difficulty on the Z-scale. The horizontal positioning of the black diamonds is related to the mean difficulty of items classified by the "ideal" judge E2 at the respective levels (A2, B1, B2 and C1). The three vertical dashed lines reflect the mean value of the mean difficulty for every two adjacent levels. In other words, these lines are exactly in the middle between two adjacent diamonds.

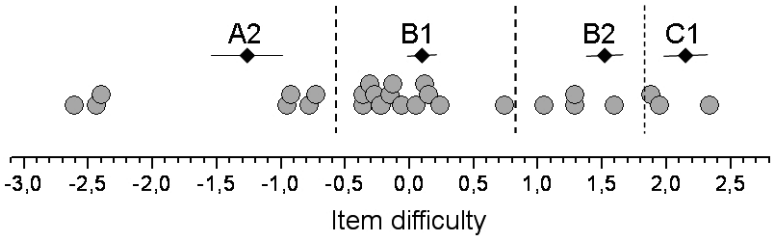


Figure 17 Cut scores for E2 (Compound Cumulative method)

The logic of setting cut scores with this method is relatively simple and is quite similar to the logic that is followed in setting cut scores with one of the examinee-centered methods, namely the method of contrasting groups (Cizek & Bunch, 2007, pp. 106-112).
According to this logic, the cut score between levels A2 and B1 is equal to the number of items (or its corresponding value on the Z-scale) which is positioned to the left of the first dashed line. Since 7 items in total are located to the left of this line, the cut score A2/B1 will be equal to 7 (in terms of the raw test score) which corresponds to -1.3 on the Z-scale (Appendix 3). In the same way, since the total number of items located to the left of the second dashed line is 20, the cut scores for B1/B2 will be equal to 20 (+1.33). Following the same logic the last cut score (B2/C1) will be equal to 24 (+2.49).
In this approach for setting cut scores, the judgments (the items assigned by the judge to ach level are used as the basis for calculating mean item difficulty for each level) as well as the empirical data (number of items with an empirical difficulty lower than the given cut score, which is calculated from the mean item difficulties of items assigned to different levels by the judge) are taken into account. In this way the main disadvantage, typical for the previous method (the Basket procedure), is overcome and one of the key recommendations for setting cut scores (Livingston & Zieky, 1982; Jaeger, 1990; Norcini, 1994; Pellegrino, J. et al, 1999; Zieky, 2001; Linn, 2003; and others), is taken into account: combining the judgments with the empirical data.
The actual data and the results from the calculations for the two judges are presented in Table 6.

Table 6 Cut scores (Compound Cumulative method)

Levels	E1					E2				
	M_X [1]	$M_{X/X+1}$ [2]	k [3]	Test score		MX	$M_{X/X+1}$	k	Test score	
				Raw score	Z-scale				Raw score	Z-scale
A2	-0.80			[0; 11]	(-4.86; -0.46]	-1.28			[0; 7]	(-4.86; -1.30]
		-0.24	**11**				-0.59	**7**		
B1	+0.33			(11; 19]	(-0.46; +1.10]	+0.10			(7; 20]	(-1.30; +1.33]
		+0.42	**19**				+0.81	**20**		
B2	+0.50			(19; 27]	(+1.10; +4.73]	+1.51			(20; 24]	(+1.33; +2.49]
		+0.56	**19**				+1.83	**24**		
C1	+0.62					+2.15			(24; 27)	(+2.49; +4.73)

1 $\mathbf{M_X}$ – mean item difficulty at level **X**
2 $\mathbf{M_{X/X+1}}$ – mean of $\mathbf{M_X}$ and $\mathbf{M_{X+1}}$
3 **k** – total number of items with difficulty ≤ **MX/X+1** (level X, according to the Compound-Cumulative method)

It is obvious that the cut scores for both judges set this way are different from the ones set with the Basket procedure. This is understandable bearing in mind that the procedures used in the two methods are different. On the other hand, however, the logical question emerges which ones of these two different sets of cut scores are the true ones. Unfortunately, this question cannot be answered. That is why the validity of the method itself for setting cut scores is of paramount importance.
Concerning the differences between the two judges, these are due to the fact that the mean difficulty of the items for the different levels is different. For instance, the mean difficulty of the items assigned to level A2 by the first judge is -0.80 while for the second it is much lower (-1.28), which explains the lower cut score for level A2/B1 for this judge. The opposite tendency can be observed for level C1 where the difference between the mean item difficulty is even greater. As a result, it appears that even if it were necessary, the cut score B2/C1 cannot be set through the judgment by E1 because it coincides with the cut score B1/B2.
The differences in the cut scores for both judges, and especially the direction of these differences, a lower first cut score (A2/B1) and a higher second cut score (B1/B2) for E2 in comparison with E1, are predictable if the way how the judgment data for E2 was obtained is taken into account. Differently from E1, the judgment E2 data does not come from a real judge. It was obtained so as to have the same frequency distribution of the items as for E1, but the classification of the items by levels was to be perfectly consistent with their empirical difficulty. With judgment data obtained in this way, it is logical that $M_{A2E1} \geq M_{A2E2}$ and $M_{C1E1} \leq M_{C1E2}$, which will lead to the corresponding differences in the cut scores.
Other reasons for such differences between the judges in real situations are the subjective nature of the performance standards and the individual strategies employed by the different judges. That is why differences in the final cut scores can be expected even if

absolute consistency with the empirical data is available for each one of the judges.
If the **advantages** of the Compound Cumulative method have to be summarized, as a whole they are the same as the advantages of the Basket procedure: **practicability**, **economy** and a **wide range of application**.
It is true that setting cut scores requires some more calculations than simply counting, all that is required by the Basket procedure. However, calculating a few descriptive statistics (mean values) cannot be regarded as complex statistical methods and would not hold back anyone who has a certificate from the upper secondary school.
The same applies to the interpretation of the results from this method – it is a bit more complex to explain to a general audience, but within the community of specialists in testing the description of the method is quite simple and easy to understand.
The main advantage of the Compound Cumulative method, compared with the Basket procedure, is that in setting cut scores both **judgments** and **empirical data** are taken into account. This leads to more adequate and justifiable cut scores and increases the validity of the method itself.
Concerning the possible misplacement of the cut scores, it is inherent in the judgment task itself. Thus it could be expected that such misplacement will be noted and the important question is how extensive it will be and whether it will be within an acceptable range. Figure 18 shows that this is the case. There are two main differences between this figure and Figure 16.
First, in the Cumulative Compound method there is no clear tendency of the misplacements depending on whether a given value of the test score is above or below the mean test score. This means that the misplacement would depend mainly on the characteristics of the frequency distribution of the difficulty of the items.
Second, while in the Basket procedure there were no even a cases of agreement (zero difference) apart from in the two extreme scores, in the Cumulative Compound method there are four more matches between the expected and the obtained cut scores.
In addition, the mean misplacement (= 1.93) is smaller than the one for the Basket procedure (= 2.29), and it is also smaller than the standard error of measurement (SEM = 2.00), which makes it acceptable according to the criterion for minimum distortion. If expressed as a percentage of the maximum possible difference, the mean misplacement is 7% which is 1,5% lower than in the Basket method.

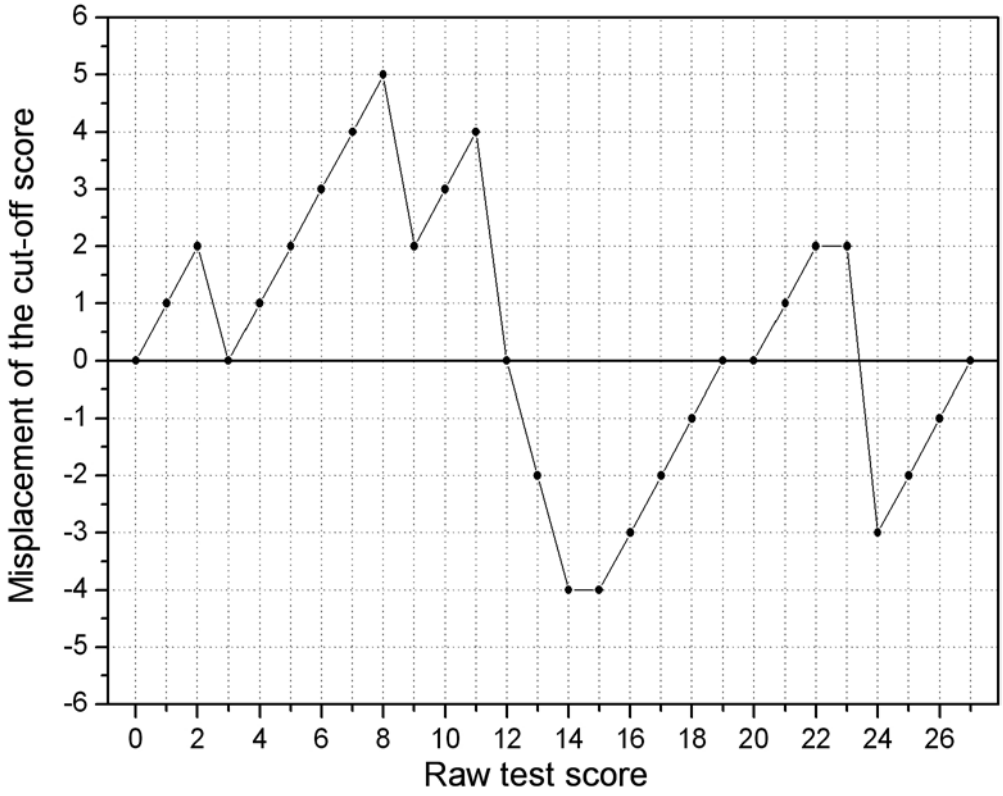


Figure 18 Cut score misplacement (expected and obtained) in the Compound Cumulative method

The main conclusion that can be drawn from this description of the Compound Cumulative method is that this is a simple method and easy to implement, with a wide range of application, which combines the judgments with the empirical data in setting cut scores. Its main disadvantage is that it allows misplacement of the judgments, but this misplacement is acceptable (< SEM)[15].

2.3.3 Cumulative Cluster method

The Cumulative Cluster method is relatively new – it was created in 2006 and has not been applied in real test situations so far. This is its first publication.
This method can be considered a modification of the Cumulative Compound method

15 Editors' note: It might be pointed out that, as used here, the need of an IRT model could be considered a disadvantage. It also needs to be studied how this method will be applicable if an IRT model, more complex than the simple Rasch model is used. Future research should explore what is its applicability if one uses the two- or three parameter logistic model. Even more noteworthy is the question what can be saved from this method if one uses only classical test theory.

because it follows the same logic. The main (and only) difference between them is the way of setting cut scores.

As the name of the method suggests, the cut scores are set by using *cluster analysis*. This is not the only method for setting cut scores using cluster analysis. Sireci's (2001) method also uses cluster analysis, but the approach and the logic are completely different.

In the Cumulative Cluster method, a cluster analysis is used applying the method of the mean distances – *'k-means cluster analysis'* (Vandev, 2003, pp. 35-39). By using it, based on the mean difficulty of the items by levels of competence for a given judgment, the items are regrouped in as many groups (clusters) as the number of the different levels of competence the judge is using in his judgment. These new groups (clusters) of items are with a pre-assigned mean difficulty equal to the mean difficulty for the separate levels according to the judgments. The thing that distinguishes the new groups from the ones defined by the judge is that the difference in the difficulty of two items from the same cluster is much smaller compared to the difficulty between two items from different clusters. In other words, by using cluster analysis the items are grouped by degree of similarity (in terms of empirical difficulty) in as many groups as used by the judge and the mean difficulty of items in each group matches the initial mean value for the levels according to the original judgment data.

Figure 19 illustrates the application of this method for setting cut scores using the judgments of the second judge. In this figure the items are classified by levels of competence, according to E2, in terms of how their horizontal position corresponds to their empirical difficulty (on the Z-scale). According to this judgment the number of items for the different levels is: A2 – 9 items; B1 – 12 items; B2 – 4 items; and C1 – 2 items.

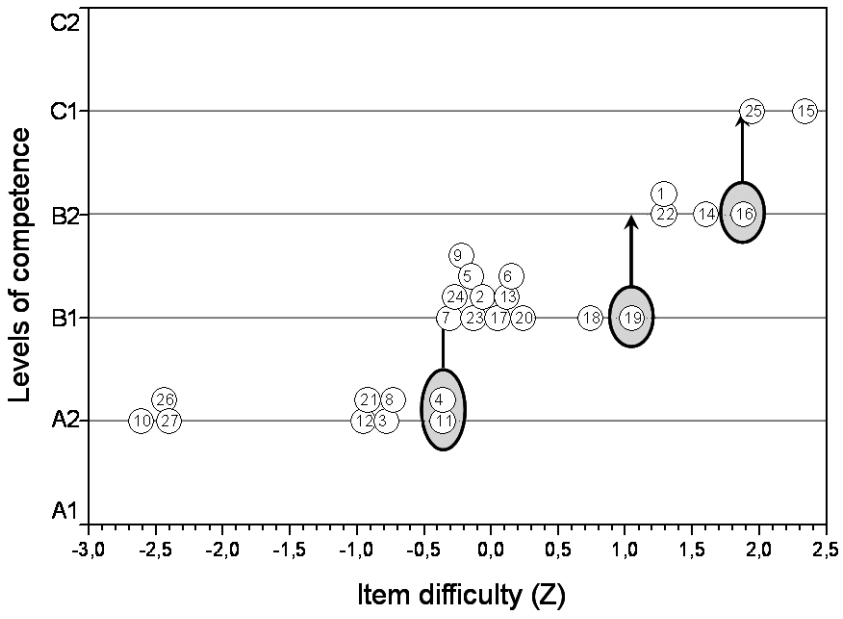


Figure 19 Cumulative Cluster method: Setting cut scores (E2)

As a result of the k-means cluster analysis, a regrouping of the items is necessary by which items № 4 and № 11 shift from level A2 to the next higher level (B1) because their difficulty is closer to the items from level B1 than for the ones at A2. For the same reason, item № 19 shifts from level B1 to level B2, and item № 16 from level B2 to level C1. As a result, the number of items at the different levels becomes: A2 – 7 items; B1 – 13 items; B2 – 4 items; and C1 – 3 items. The setting of the cut scores is implemented in the same way as in the Basket procedure – each cut score is equal to the total number of items which belong to the levels of competence under the respective cut scores. In this case the cut scores are as follows: A2/B1 = 7; B1/B2 = 20; B2/C1 = 24.
If we compare these cut scores with the ones from the same judge in the Compound Cumulative method, we will see that they are identical, which supports the similarity between the two methods. The same conclusion can be drawn for the first judge as well if we look at Table 7. This, of course, does not mean that the cut scores obtained by these two methods will always be the same. However, it can be expected that they will be close to each other.

Table 7 Cut scores (Cumulative Cluster method)

Levels	E1					E2				
	M_x [1]	k_x [2]	c_x [3]	Test score		M_x	k_x	c_x	Test score	
				Raw score	Z-scale				Raw score	Z-scale
A2	-0.80	9		[0; 11]	(-4.86; -0.46]	-1.28	9		[0; 7]	(-4.86; -1.30]
			11					**7**		
B1	+0.33	12		(11; 19]	(-0.46; +1.10]	+0.10	12		(7; 20]	(-1.30; +1.33]
			19					**20**		
B2	+0.50	4		(19; 27]	(+1.10; +4.73]	+1.51	4		(20; 24]	(+1.33; +2.49]
			19					**24**		
C1	+0.62	2				+2.15	2		(24; 27)	(+2.49; +4.73]

1 $\mathbf{M_x}$ – mean item difficulty at level **X**
2 $\mathbf{k_x}$ – number of items at level **X** according to the judgment data
3 $\mathbf{c_x}$ – number of items at level **X** according to the cluster analysis

The same can be argued for the degree of misplacement of the cut scores. Only for one value of the raw test score is a difference between the misplacement in the Cumulative Compound and Cumulative Cluster method apparent and this is the value of 12. For this value the misplacement in the Cumulative Cluster method is -2 points, i.e. the calculated value is 14, while in the Compound Cumulative method the obtained cut score matches the expected one.
In the end this leads to increasing the mean value of the misplacement, which becomes equal to the standard error of the measurement (= 2.0), and the number of matches between the expected and obtained cut scores decreases to three.
If the main characteristics of the Cumulative Cluster method have to be summarized, its major advantages, as in the previous two methods, are practicability, economy and wide range of application. Its additional advantage is that in setting cut scores both the

judgments and empirical data are used, which is one of the necessary prerequisites for the validity of a given method.
Its main disadvantage is again the eventual misplacement of the obtained cut scores in relation to the expected scores, but this misplacement, although possible, is within an acceptable range (≤ SEM).
What distinguishes it from the preceding two methods is that, in setting cut scores, a relatively more complex method is used (cluster analysis) which requires appropriate software. That is why the method is more appropriate for scientific research and external validation of cut scores that were set using other methods.
If we compare this method with the preceding ones, we could say that, in terms of validity, the Cumulative Cluster method is preferable to the Basket procedure, but not preferable to the Compound Cumulative method because of its statistical complexity.

2.3.4 ROC-curve method

The ROC-curve method is a relatively new method as well. It was developed in 2006, not applied in real test situations until now and this is its first publication. Two requirements were taken into account in its development:
First, the method should have a wide range of application, i.e. it should be possible to apply it to tests regardless of the theoretical foundation for their development and the test item format.
Second, in setting cut scores the judgments and the empirical data should both be taken into account.
In this method, as its name shows, the *receiver operating characteristic curve* (or shortly – *ROC*) is used.
Such curves, which come from the signal detection theory, are applied successfully in making classification decisions in many areas of science and are very popular in medical diagnostics. In the field of achievement testing and setting cut scores, they are still not very widely used. Among the imposing number (over 8000) of scientific publications (see Figure 1) in this area, there are only two (Wiliam, 1998; Hintze & Silberglitt, 2005) which discuss the application of the ROC-curves in setting cut scores.
The application of the ROC-curves for setting cut scores suggests transformation from the ordinal scale of the judgments into a dichotomous classification of the test items by levels of competence (Appendix 2). According to this dichotomization, for a given level of competence ***X***, there are two options for each item: a **true/false** answer by an examinee who is at a given level. The first possibility (**correct** answer) will apply when, according to the judge, this item belongs to level ≤ ***X***. The second possibility (**incorrect** response) will apply in cases when, according to the judge, this item can only be answered correctly at level > ***X***.
Figure 20 is an illustration of the dichotomization for level A2 for the first judge (E1). According to his judgment, only nine of the items can be answered correctly by examinees whose level of competence is A2 and they are located in the lower row, corresponding to their empirical difficulty. All items which, according to the judge, can be answered correctly at level > A2 but cannot be answered correctly at level A2, are in the upper row (again corresponding to their difficulty).

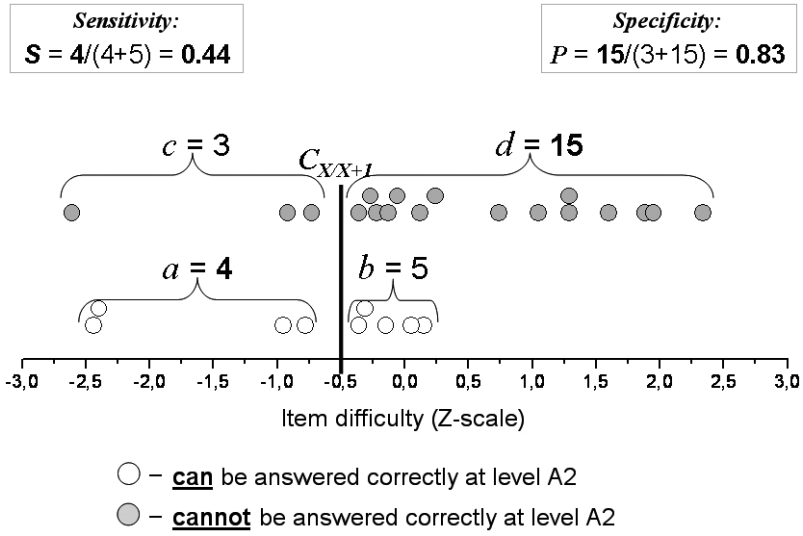


Figure 20 Dichotomization of the judgment (E1) for level A2

If we accept for the sake of discussion that the cut score **A2/B1** is equal to -0.5 (the thick vertical line – $C_{X/X+1}$), then this cut score will divide the items into four groups:

a items classified **correctly** by the judge as **correctly** answered at this level;

b items classified **incorrectly** by the judge as **correctly** answered because their difficulty is higher than the respective cut score ($C_{X/X+1}$);

c items classified **incorrectly** by the judge as **incorrectly** answered because their difficulty is lower than the respective cut score ($C_{X/X+1}$);

d items classified **correctly** by the judge as **incorrectly** answered because their difficulty is higher than the respective cut score ($C_{X/X+1}$).

The analysis of the ROC-curves is closely linked to two key terms, namely sensitivity and specificity, which in turn are related to the two types of correct classification (***a*** and ***d***). In the particular case, these two terms can be defined as follows:

- **Sensitivity (*S*)** is the proportion of the items with difficulty < $C_{X/X+1}$ which, according to the judge, can be answered correctly by an examinee at level ***X*** and can be calculated using the formula: ***S*** = *a/(a+b)*.
- **Specificity (*P*)** is the proportion of the items with difficulty > $C_{X/X+1}$ which, according to the judge, can be answered correctly by an examinee at level ***X*** and can be calculated using the formula: ***P*** = *d/(c+d)*.

These indices for sensitivity and specificity (*S* and *P*) can vary within the interval [0; 1] and obviously depend on the respective cut score, i.e. for different cut scores they will be different. For example, for the cut score = -0.5, sensitivity will be equal to 0.44, and specificity to 0.83. The question that logically arises is at what cut score the two indices will take maximum possible values.

To answer this question, it is obvious that the values of these indices for each possible cut scores have to be calculated (Appendix 5). After the values for sensitivity and specificity

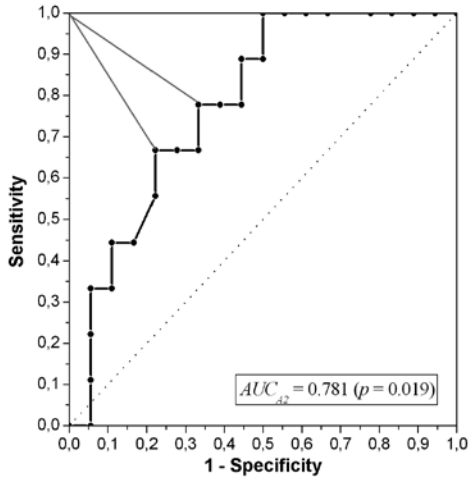
indices are known, the cut score setting will be reduced to solving the classical optimization problem, whose purpose is to find that point on the Z-scale for which the number of the correct classifications is maximum and, respectively, the number of the incorrect classifications will be minimum.

In the particular case, since the test consists of 27 items, the number of possible cut scores should be 28 (one between every two items, plus two more at both ends of the interval). However, since there are two pairs of items with the same difficulty, the number of the maximum possible cut scores will be 26.

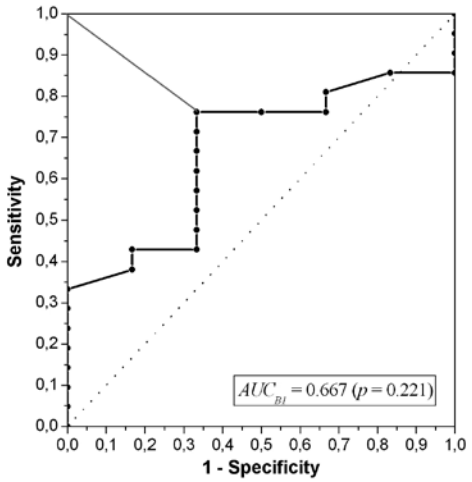
If we plot the points with their coordinates (S_C and 1-P_C) on the axes and join these points with straight lines, we will get the ROC-curve for a given judgment at the respective level of competence. Figure 21 shows the ROC-curves for three levels of competence (A2, B1, B2) for the first judge (E1) and for level A2 for the second one (E2).

The smoothness of the curve depends on the number of points, which in the concrete case will depend on the number of test items in each particular test. Here this number is rather small and leads to quite "choppy" curves.

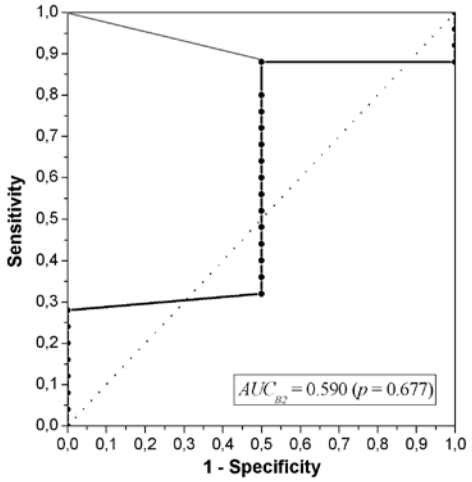
The ROC-curve is the main **criterion for quality** when setting cut scores with this method in the achievement tests. The dashed diagonal line in each of the four figures shows how the ROC-curve would look if the judge had classified them by random guessing (for example tossing a coin). Therefore, the higher from the diagonal the ROC-curve is located, the better (i.e. different from random guessing) are the judgments. The judgments of the second judge are perfect in this case because the ROC-curve coincides with the two axes for sensitivity and specificity.



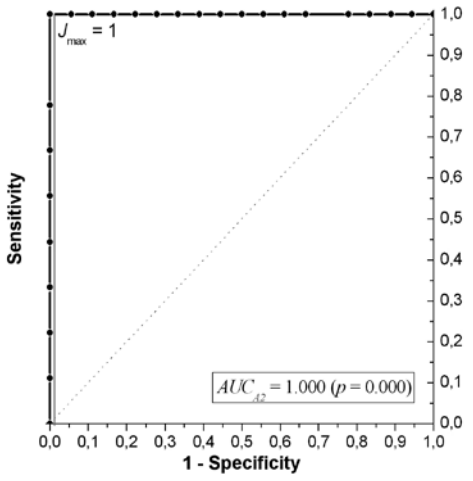
(a) E1 – level A2



(b) E1 – level B1



(c) E1 – level B2



(d) E2 – level A2

Figure 21 ROC-curves

The most frequently used numerical index for accounting for the quality of the judgment is **the area under the curve** (AUC). Since the area under the diagonal is equal to 0.5, if a judgment is better than random guessing, it is necessary for its ***AUC*** index to be significantly higher than 0.5. This condition is satisfied only for level A2 for the first judge E1, the area under the curve is statistically different (p = 0.019) from 0.5, i.e. distinct from random guessing.

This confirms one of the conclusions drawn in the analysis of the consistency between the judgments and the empirical data, namely that the first judge is better in classifying the

items at the lower levels.
In the case of the second judge, whose judgment was made with perfect consistency with the empirical data, it can be seen that his ROC-curve for level A2 is perfect, with a maximum area under the curve (***AUC*** = 1). The ROC-curves for the rest of the levels of competence for the second judge are analogous to the one shown in Figure 21-d and, for this reason, are not presented here.
From the preceding discussion, it is clear that the ROC-curve method provides an additional opportunity for an in-depth analysis of the consistency between the judgment and the empirical data and for a comparison between several judgments.
The main questions, however, namely, what is the optimal cut score and how to set it remain unanswered. Intuitively it is clear that the cut score has to be such that its point on the curve should be as distant as possible from the diagonal and as close as possible to the upper-left corner of the graph.
This is because the further from the diagonal (in the direction to the upper-left corner) a given point is, the less the probability for random guessing, i.e. the judge does not rely on chance but makes his or her judgments on the basis of his or her experience and good knowledge about the behaviour of the test items.
On the other hand, the upper-left corner of the graph has the co-ordinates (0; 1), which means that for the judge both indices (sensitivity and specificity) have a maximum value equal to one. In other words, if for a given cut score the respective point on the ROC-curve matches with the upper-left corner (Figure 21-d), the classification for this cut score is perfect, meaning that there are no incorrect judgments for it.
These intuitive and rather geometric considerations correspond to the two most often used criteria for setting optimal cut scores (Perkins & Schisterman, 2006) in the case when the potential consequence of the incorrect judgments is the same.
The first optimization criterion (***D_{01}***) is the so called **criterion for maximum proximity to a point**. The algebraic expression of this criterion is: $\sqrt{(1-S_C)^2+(1-P_C)^2}$, which is an algebraic expression of the distance between the upper-left corner and the point of the ROC-curve that corresponds to the cut score (***C***). The index ***D_{01}*** can take values between 0 and 1, and according to the criterion, the optimal cut score is the one for which ***D_{01}*** has the lowest value.
On each of the three ROC-curves for E1 (Figure 21 – a, b, c), the points for which ***D_{01}*** has a minimum value are connected with the upper-left corner by a thick grey line and in Appendix 5 the exact values for ***D_{01}*** are provided for each possible cut score at levels A2, B1 and B2 according to the judgments of the first judge.
As can be seen in the graphs in Figure 21-b, for levels B1 and B2, a single solution exists for the optimization problem and the cut scores matching the points that are closest to (0; 1), expressed in terms of the number of correctly answered items, are respectively: B1/B2 = 18 and B2/C1 = 23 (Appendix 5 and Table 8). For level A2, however, there is no single solution because there are two points on the curve which are at an equally close (0; 1), which is the minimum distance for all possible cut scores. This impedes the choice of a cut score for A2/B1 and imposes the use of another criterion.
The second optimization criterion is based on the so called **Youden index (*J*)**, which is calculated using the formula: ***$J_C = S_C + P_C - 1$***. According to this criterion, the optimal cut

score is the one for which the value of J_c is maximum. Using school geometry, it can easily be proved that the value of the Youden index is equal to the length of the vertical lines connecting a given point of the ROC-curve with the diagonal. The absolute maximum of this index is 1 and is reached at the point (0; 1).
Besides its use as an optimization criterion, the Youden index is also used as a criterion for the quality of the judgment as a whole or for a particular cut score. Appendix 5 contains the values for this index for the first judge for each possible cut score at levels A2, B1 and B2. As can be seen, the value of this index does not exceed 0.5 for any possible cut score. This maximum value (0.5) is quite low and indicates that the given judge does not adequately judge the item difficulty. Unfortunately, even in real test situations, the number of the judgments with a similar and lower quality is often not insignificant, which puts the adequacy of the cut scores set in such cases in serious doubt, Concerning the use of Youden index as an optimization criterion, Appendix 5 shows that the cut score for which it reaches its maximum at levels B1 and B2 coincides with the optimum cut score according to the first criterion.
For level A2, however, such correspondence does not exist. Moreover, according to the Youden criterion, the optimal cut score for level A2 coincides with the optimal one for level B1. Given this situation, it is preferable to choose one of the two cut scores which are optimal according to the first criterion. Keeping in mind that the cut score A2/B1 is the lowest of the three cut scores (A2/B1, B1/B2 and B2/C1), it would be more acceptable to choose the lower one of the two, namely the point which corresponds to 10 correctly answered items (Table 8).

Table 8 Cut scores (ROC-curve method)

Levels	E1					E2				
	M_x[1]	k_x[2]	c_x[3]	Test score		M_x	k_x	c_x	Test score	
				Raw score	Z-scale				Raw score	Z-scale
A2	-0.80	9		[0; 10]	(-4.86; -0.65]	-1.28	9		[0; 9]	(-4.86; -0.85]
			10					**9**		
B1	+0.33	12		(11; 18]	(-0.65; +0.89]	+0.10	12		(10; 21]	(-0.85; +1.57]
			18					**21**		
B2	+0.50	4		(19; 23]	(+0.89; +2.14]	+1.51	4		(22; 25]	(+1.57; +2.92]
			23					**25**		
C1	+0.62	2		(24; 27]	(+2.14; +4.73]	+2.15	2		(26; 27)	(+2.92; +4.73)

1 MX – mean item difficulty at level X
2 kX – number of items at level X according to the judgment data
3 cX – number of items at level X according to the ROC-curve method

On an intuitive level, both optimization criteria are close to each other because both are striving to maximize the proportion of the correct classifications at the expense of incorrect ones. In practice, however, as can be seen from the results at level A2, their application does not always lead to the same optimal cut scores. Which one should be used in particular test situations is a question that cannot be answered in a straightforward

manner although the preference is towards the Youden index (Perkins & Schisterman, 2006). The reasons for this are its easier calculation, its clearer interpretation and the wider currency it has gained.
If we compare the cut scores set using the ROC-curve method (Table 8) with the ones obtained using the previous methods, we will notice that the cut scores for the second judge coincide with the ones set using the Basket procedure. This is a result that can be foreseen, having in mind the way the judgments were obtained – with perfect correspondence with the empirical data.
For the first judge, however, there is no such correspondence between the cut scores set by the Basket procedure and the ROC-curve procedure. The main reason for this difference is that in using the ROC-curve method for setting cut scores, not only the judgment are taken into account but the empirical data as well. The bigger the difference between the subjective judgments and the objective results, the bigger will be the difference between the cut scores obtained using different methods.
As a whole, however, the differences between the cut scores are larger between the Basket procedure and the ROC-curve than between the ROC-curve and the other two methods.
Concerning the possible misplacement of the cut scores, it is present in this method too, as expected, and this misplacement matches fully the one in the Basket procedure (Figure 16) and the mean misplacement (2.29) is higher than the standard error of the measurement. In other words from the point of view of the potential misplacement, the Compound Cumulative and the Cumulative Cluster methods are preferable to the ROC-curve method.
The main advantages of the ROC-curve method are practicability, economy, wide range of application and consistency of the judgment with the empirical data. An additional advantage is that it provides an opportunity for an in-depth analysis of the consistency between the judgments and the empirical data, which is not possible in the preceding methods.
A disadvantage of this method (along with the main disadvantage – misplacement of the cut scores) is the necessity of using more complex statistical methods and corresponding software in setting the cut scores. The complexity of the statistical method leads to consequent difficulties in describing it, and in gaining acceptance of the method from a broader audience (examinees, parents, teachers, etc.), which might limit its application. That is why the ROC-curve method is more appropriate as a secondary method – for scientific research and the validation of already set cut scores – than as main method for setting cut scores.

2.3.5 Item Mastery method

From a chronological point of view, the Item Mastery method (Verhelst & Kaftandjieva, 1999; Kaftandjieva, & Verhelst, 2000; Kaftandjieva, et al, 2000) appeared first among all the methods for setting cut scores that have been presented so far. The method was created and applied for the first time during the first phase of the international European project for Internet-based testing – DIALANG (http://www.dialang.org/english/index.htm) in 1999. Subsequently the method was applied in several more projects in Finland in the area of the foreign language testing.
As with the preceding methods, the Item Mastery method is applicable to all kinds of test

items regardless of their weight or method of scoring. However, this method is applicable only for tests that were developed using IRT.
The Item Mastery method also requires preliminary dichotomization of the judgments for each level of competence (Appendix 2) and in this it resembles the ROC-curve method. The necessity of dichotomization is easy to see, bearing in mind that the judgment in the initial version of the method was multi-stage, as in each stage the judges had to classify the test items on the basis of their answer to the following question: Do you think (yes/no) that an examinee with language competence at level X will answer this item correctly? The number of stages (rounds) in this judgment depends on the number of cut scores that we want to set. If they are three, as in this particular case (A2/B1, B1/B2 and B2/C1), then the number of rounds would be three as well, and the levels for which the question would be addressed concerning the judgment would be respectively A2, B1 and B2.
This formulation of the judgment task allows for greater control over the reliability of the judges in the sense of consistency of their judgments for the separate rounds. On the other hand, however, the procedure takes more time, and due to considerations concerning practicability and economy, the judgment procedure was replaced with the current format of the judgment.
The other similarity between the two methods (ROC-curve and Item Mastery) is that, in both methods, optimization of the decision-making is sought and in this sense the cut score being set is optimal. The difference is in the criterion of this optimum.
The algebraic formula for the **optimum criterion** in the Item Mastery method is as follows:

$$\min_{x} L(x) = \sum_{i=1}^{k} (I_{x \le Zi} \cdot D_i \cdot (Z_i - x)^2 + I_{x > Zi} \cdot (1 - D_i) \cdot (Z_i - x)^2)$$

where

$L(x)$	loss function
k	number of test items
Z_i	difficulty for item i in Z-scale
D_i	result of dichotomizing item i for a given judge (**0** or **1** depending on whether the item can or cannot be answered correctly by a given examinee at a given level of competence according to the judge)
$I_{x \le Zi}$	dichotomous variable equal to **1** at $x \le Zi$ and equal to **0** otherwise
$I_{x > Zi}$	dichotomous variable equal to **1** at $x > Zi$ and equal to **0** otherwise

In practice, the application of this criterion leads to finding the point on the Z-scale for which the summary quadratic distance to items classified incorrectly by the judge is minimal.
The hypothetical example shown in Figure 22 illustrates the application of this criterion for setting the optimal cut score. In this hypothetic example, two cut scores are presented – X and Y. The number of incorrect judgments, three, is the same for each cut score (the items marked with ⊗ in Figure 22). If the ROC-curve method is applied to this example, the values for both optimum criteria would be the same for the points X and Y. If the Item Mastery

method is applied, however, the point Y would be preferable for the cut score rather than point X because the summary distance from it to the three incorrect judgments is smaller than for X. This summary distance, presented in the upper-right corner of the two graphs in Figure 22 and expressed on the Z-scale, is over two units for point X and below two units for point Y.

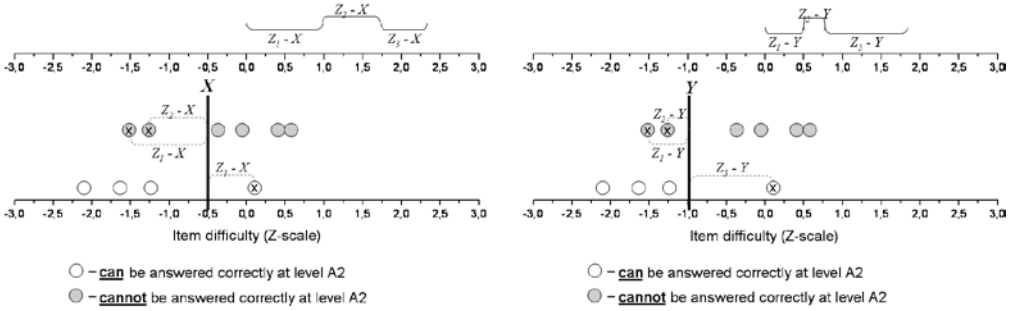


Figure 22 Item Mastery method – a hypothetical example

The main difference between optimum criteria in both methods is that while in the ROC-curve method it is about the number (or proportion) of items, in the Item Mastery method it is about minimal distance. The distance as a criterion makes it impossible to apply the method to tests developed using classical test theory. The distance between the two objects (in this case item and cut score) can be calculated only if the coordinates for both of them are expressed using the same scale. Since, in classical test theory the item difficulty and the degree of the measured ability are expressed on different scales, the Item Mastery method cannot be applied.

The other difference between the two methods is that, as can be proved (Verhelst & Kaftandjieva, 1999), the function of the potential loss – the *loss function* (1) – has a minimum. In cases when the pattern of judgment is different from the perfect pattern (Guttman pattern or absolute consistency with the empirical data), this minimum is unique. The point x0 on the Z-scale, for which this minimum is reached, is the optimal cut score in the sense of the optimum criterion used in this method. The value of the function itself $L(x_0)$ can be used as an index of the quality of the judgment of this judge for a given level. In this way the judgments from a single judge for different levels or at different levels or even the judgments from different judges can be compared. Moreover, the loss function can be calculated separately for each item with a given fixed cut score. This allows analysing the influence of different factors related to the content and the format of the test items on how consistent the judgment data are with the empirical data.

The results presented in Figure 23 confirm the conclusions drawn from the analysis of the consistency, according to which items № 11 and № 10 are the ones that appeared to be the most problematic for the first judge. The loss function has, in fact, the highest values for these items. The other conclusion that can be drawn is that while, for item № 11, the judge meets difficulties in defining its status for levels B1 and B2, for item № 10 the problems arise at level A2.

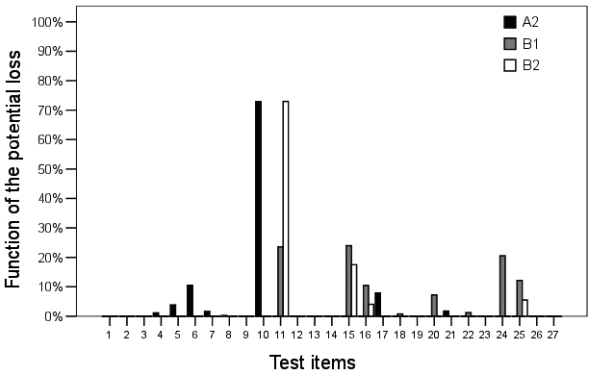


Figure 23 Degree of adequacy of the judgment (E1) for the separate test items

If the dichotomization of a given judgment is perfect, i.e. fully consistent with the empirical difficulty of the items, the loss function will reach its absolute minimum ($L(x) = 0$), but that minimum is not unique. It will be reached for all points of the interval of the Z-scale between the Z-value of the most difficult item that is mastered and the Z-value of the easiest item that is not mastered. In this case, the choice of a point from this interval, which can be set as the cut score, is subjective.
The cut scores between the separate levels for this particular case (E1 and E2) are presented in Table 9. They, as the values of *L(x)*, used in constructing Figure 23, are calculated using a computer program developed especially for this purpose (cutoffsq.exe) by Norman Verhelst, who is a co-author of the Item Mastery method.

Table 9 Cut scores (Item Mastery method)

Levels	E1					E2				
	M_x[1]	k_x[2]	c_x[3]	Test score		M_x	k_x	c_x	Test score	
				Raw test score	Z-scale				Raw test score	Z-scale
A2	-0.80	9		[0; 10]	(-4.86; -0.65]	-1.28	9		[0; 11]	(-4.86; -0.46]
			10					**11**		
B1	+0.33	12		(11; 18]	(-0.65; +0.89]	+0.10	12		(11; 19]	(-0.46; +1.10]
			18					**19**		
B2	+0.50	4		(19; 20]	(+0.89; +1.33]	+1.51	4		(20; 22]	(+1.10; +1.84]
			20					**22**		
C1	+0.62	2		(21; 27]	(+1.33; +4.73]	+2.15	2		(23; 27)	(+1.84; +4.73)

1 $\mathbf{M_x}$ – mean item difficulty at level **X**
2 $\mathbf{k_x}$ – number of items at level **X** according to the judgment data
3 $\mathbf{c_x}$ – number of items at level **X** according to the Item Mastery method

The cut scores set in this way by the first judge match perfectly with the ones set by the ROC-curve method. Concerning the cut scores of the second judge, they match none of the ones set with the preceding four methods. Moreover, the cut score A2/B1 = 11 is higher

than in any of the other methods, and the other two cut scores are lower than in all the other methods. Since we are talking about the second judge (whose judgment is perfect from the point of view of consistency with the empirical data), the question that logically arises is which of all cut scores is, if not the true one, at least the most adequate. This question, although it has no answer, draws attention to the criteria for minimum distortion of the judgments.

The difference between the expected and the obtained cut scores obtained with the Item Mastery method is presented in Figure 24. As the figure shows, besides the two extreme cases (0 and 27), there are nine more cases where no misplacement is present. In other words, the Item Mastery method gives the opportunity for a precise prediction of 11 out of 28 cut scores (39%) – a result achieved by none of the preceding methods. The mean misplacement in absolute values (= 0.96) is also significantly lower (more than two times) compared to the other methods and is ≤ ½.*SEM*.

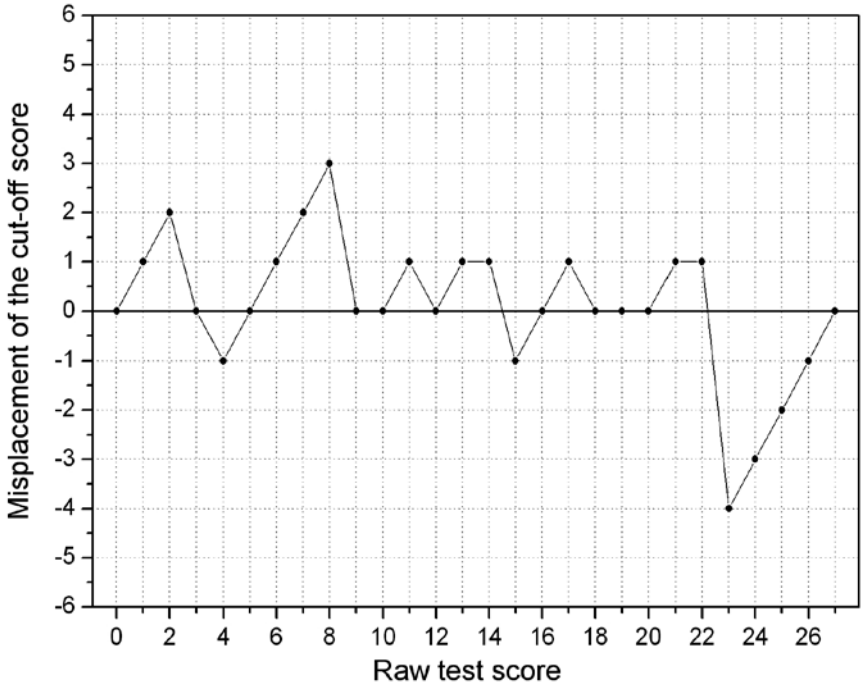


Figure 24 Misplacement of the cut score (expected and obtained) in the Item Mastery method

In summary, the main **advantages** of the Item Mastery method are:

- Practicability and economy of the judgment;
- Applicability for test items regardless of their format, weight and scoring scheme;
- Integration of the judgment and the empirical data in the final setting of the cut scores;
- Possibility for an analysis of the adequacy of the judgment for the separate levels as a whole, and also for each item separately;
- Minimum degree, in advance, of misplacement of the cut scores from the expected.

The main **limitation** of the method is that it is applicable only to tests developed using IRT. Another disadvantage is the relatively more complex statistical method that is required in setting the cut scores. This, on the one hand, requires additional software, and on the other, makes it harder to present the results to a broad audience.
In summary, the Item Mastery method is more appropriate in big assessment projects, mostly in the external assessments where the IRT is common practice and the analysis of the data is usually implemented by qualified specialists – psychometricians and statisticians. For the needs of the internal assessment and tests developed for formative assessment the other methods (Cumulative Compound and Basket procedure) are recommended.

2.3.6 Level Characteristic Curve method

The Level Characteristic Curve method was developed in 2004 and was applied to setting cut scores in several projects in the area of foreign language testing in Spain, but this is its first publication.
As the name of the method suggests, it is based on the Level Characteristic Curves for the different levels of competence, defined by the respective judgment data. As the Level Characteristic Curve is a term from IRT, the method is applicable only for tests developed using IRT[16].
The Level Characteristic Curve of a given test item shows the relationship between the degree of development of the measured characteristic and the probability for a correct answer to the item. The formula used to estimate this functional relationship is different depending on the different probability models. In the Rasch model, which is used in this illustrative example, the corresponding formula is: $P(x) = (1 + e^{-(x-Z_i)})^{-1}$, where Z_i is the difficulty for the respective item. Figure 25 shows the Level Characteristic Curve for the first test item, whose difficulty (Z_1) is equal to +1.286.

16 Editors' note: The term 'Level Characteristic Curve' is not standard usage in IRT. The older term used is 'item characteristic curve' while in modern literature the term 'item response function' is the preferred one. However, the usage of the term seems to be common in ROC-curve literature.

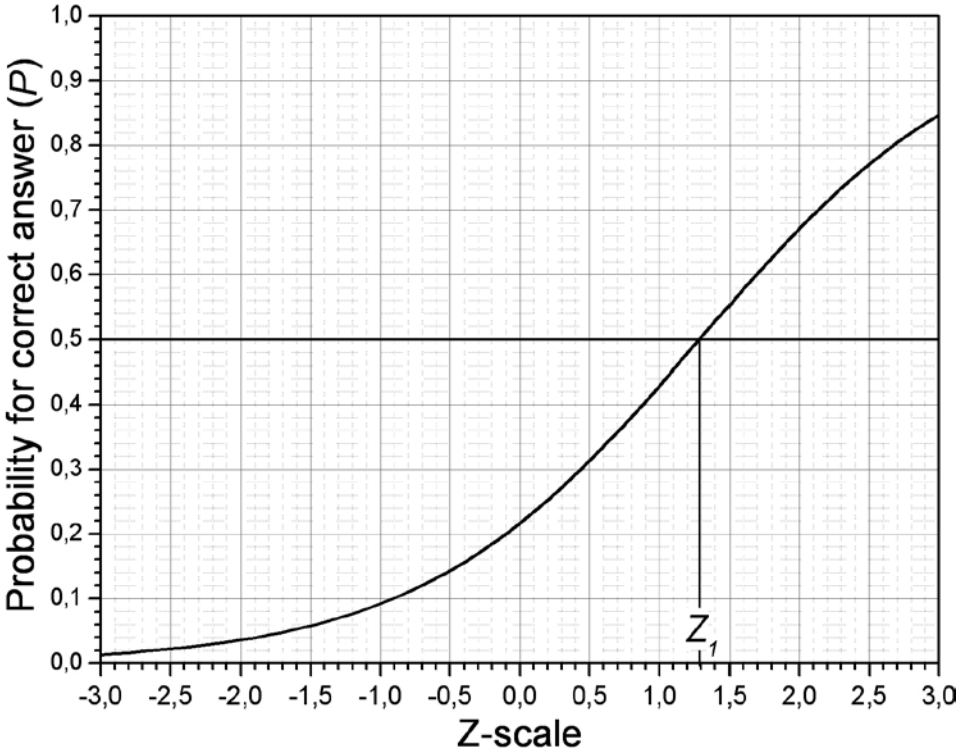


Figure 25 Level Characteristic Curve for test item № 1

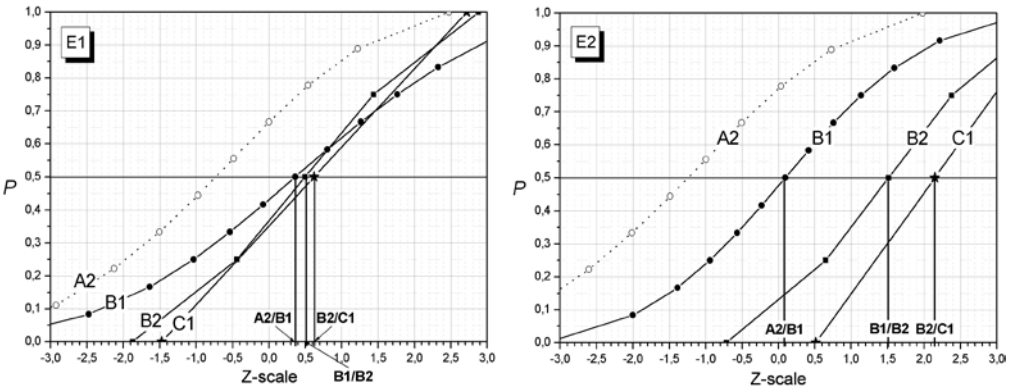
As can be seen in Figure 25, the curve increases monotonously, i.e. with an increase of the measured characteristic (in this case – reading comprehension) the probability for a correct answer also increases. Moreover, for all examinees whose test score is higher than Z_1, the probability to provide a correct answer is higher than 0.5 (or 50%).
Such a Level Characteristic Curve can be drawn not only for each single item but also for the entire test or for a group of items – for example the items which, according to given judgment, are at level *X*. In such a case, the probability for a correct answer to these items will be higher than 50% for each examinee with a test score higher than Z_x (the point on the Z-scale for which $P = 0.5$). In other words, the point Z_x is the border (cut-off) value between the level *X* and its preceding level (*X*–1).
This is the logic of the Level Characteristic Curves[17]. According to this method, the cut score

17 Editors' note: As the explanation is brief and may not be easy to follow, we wish to add a few remarks. Adding the item response curves for all items in the test produces what is known as the test characteristic curve or function. Its value runs from zero to the number of items. Dividing by the number of items brings the range of the function back to the (0,1) interval. Instead of taking the sum across all items, it seems that the author takes the sum of all the curves being classified into the same level (and then takes the average). In this sense, the reference to such a curve as Level Characteristic Curve makes sense.

between two consecutive levels of competence (*X*–1 and *X*) is the point on the Z-scale for which the Level Characteristic Curve of the items at level *X* crosses the straight line *P* = 0.5. This means that in setting a given cut score using this method, the only information that is used is the data for the test items which, according to the judgment, belong to the **next level** (the level which is immediately **above** the respective cut score). This distinguishes this method from the preceding ones and especially from the Basket procedure, in which the cut score is set on the basis of those items which the judge had classified as belonging to the levels which are **below** the respective cut score.

In its logic the Level Characteristic Curve method is to some extent getting close to hierarchical IRT modelling (Janssen, 2000), the method for defining achievement levels using IRT-estimated domain scores (Schulz et al., 1999), the cognitive component model (McGinty & Neel, 1996) and the attribute hierarchy model (Sadesky & Gushta, 2004). The main differences between them are in choosing the concrete theoretical model, the nature of the judgment task and the classification schemes used in drawing the Level Characteristic Curves.

The Level Characteristic Curves for the separate levels of competence, corresponding to the two judgments (E1 and E2) are presented in Figure 26, and the corresponding cut scores (on the Z-scale and raw test score) are presented in Table 10.



(a) (b)

Figure 26 Level Characteristic Curves and cut scores for E1 and E2

In a qualitative judgment which is congruent with the empirical data, it could be expected that the Level Characteristic Curves for the separate levels will not cross and will be clearly distinguishable from each other as the curve for each next level will be to the right of the preceding one. This can be seen for the second judge (E2), whose judgment were derived to be in a perfect agreement with the empirical data.

The first judgment (Figure 26-a) does not meet this expectation because the curves for levels B1, B2 and C1 are located very close to each other and cross. The only positive characteristic of these judgments is that the obtained cut scores between the separate levels also increase as the level increases: $C_{A2/B1} < C_{B1/B2} < C_{B2/C1}$. Although it may sound

trivial, this requirement is important and unfortunately in real situations for some judges it is not true.

Table 10 Cut scores (Level Characteristic Curves method)

Levels	E1					E2				
	M_x[1]	k_x[2]	c_x[3]	Test score		M_x	k_x	c_x	Test score	
				Raw test score	Z-scale				Raw test score	Z-scale
A2	-0.80	9		[0; 15]	(-4.86; +0.36]	-1.28	9		[0; 13]	(-4.86; +0.09]
			15					**13**		
B1	+0.33	12		(15; 16]	(+0.36; +0.50]	+0.10	12		(13; 20]	(+0.09; +1.51]
			16					**20**		
B2	+0.50	4		(16; 17]	(+0.50; +0.62]	+1.51	4		(21; 23]	(+1.51; +2.15]
			17					**23**		
C1	+0.62	2		(17; 27]	(+0.62; +4.73]	+2.15	2		(24; 27)	(+2.15; +4.73)

1 $\mathbf{M_x}$ – mean item difficulty at level X
2 $\mathbf{k_x}$ – number of items at level X according to the judgment data
3 $\mathbf{c_x}$ – number of items at level X according to the Level Characteristic Curve method

These three cut scores for E1 (Table 10) are, however, so close to each other that the difference between them is smaller than the standard error of measurement (*SEM* = 2), which in practice makes them indistinguishable from each other.
The reason for this proximity between the cut scores is the small differences in the mean item difficulties for levels B1, B2 and C1 (Table 10: E1 – column M_x), which leads to the curves being close to each other. It is like that because the degree of consistency between the judgment and the empirical data is not high enough and the judge has classified difficult items at lower levels and easier items at higher levels. The logical result is that the cut scores cluster in the middle of the interval because in the Level Characteristic Curves method, as well as in the Basket procedure, the cut score are derived from the judgments only, without making adjustments by taking into account the empirical difficulty of the items.
Actually the only level that is clearly distinguishable from the others is level A2. In this method, however, this level is not taken into account in setting the cut scores A2/B1, B1/B2 and B2/C1, since all of them are above it.
A more thorough analysis of the data from Table 10 (column M_x) shows that, except for constructing the Level Characteristic Curves, there is another, much faster and easier (although mechanical) procedure for setting the cut scores. According to this procedure the cut score $\boldsymbol{C_{X\text{-}1/X}}$ is equal to the mean of the difficulty of items which are at level ***X***, according to a given judge.
A comparative analysis of the cut scores for E1 obtained by applying the different methods shows significant differences especially concerning the first cut score (A2/B1), which is higher by several points (between 4 and 6) in the Level Characteristic Curve method than it is in the other methods. The reason for this difference is that in setting this cut score only the items which, according to the judge, are at level B1 are taken into account, but not the

ones at level A2. On the other hand, as the analysis of the consistency of judgments shows, the judgment of this judge is most adequate at this particular level (A2) while at the other levels it is more or less problematic.
Actually the tendency to arrive at a higher A2/B1 cut score with the Level Characteristic Curve method is also observed in the cut scores derived from the judgements by E2, although the differences there are comparatively smaller.
The difference between the cut scores obtained using different methods, with a maximum of 6 points (= 3x*SEM*), again brings up the question which one of these cut scores is the most adequate and whether this difference is due to a potential misplacement of the cut scores. In other words, for each of these methods, when a given judge has understood and correctly interpreted the judgment task, knows in advance the cut score that he or she wants to set, and has access to all the empirical data of the items, will there be any misplacement in the obtained cut scores, and if there is, in what direction will it be?

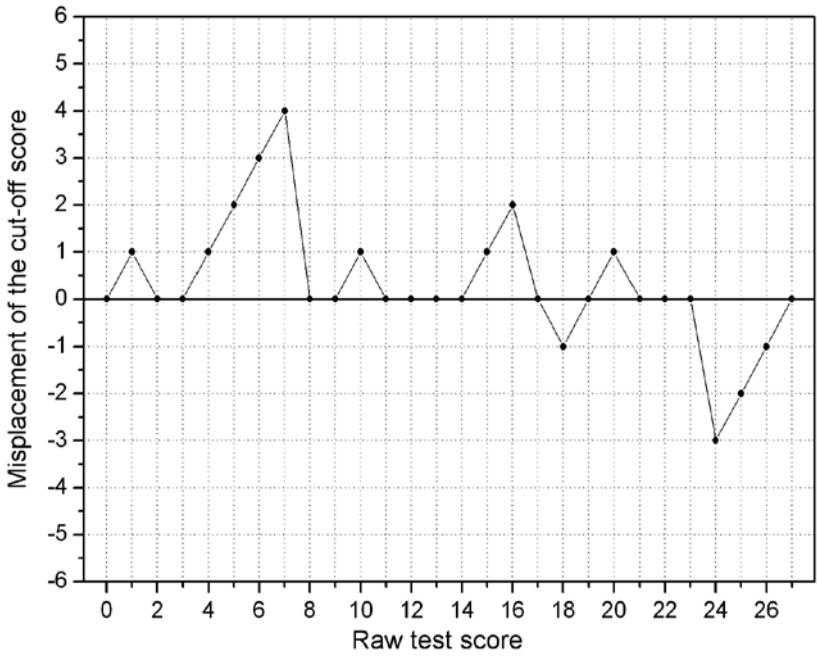


Figure 27 Misplacement of the cut score (expected and obtained)in the Level Characteristic Curves method

Figure 27 shows what the difference between the expected and obtained cut scores is for the Level Characteristic Curve in this case. As can be seen in this figure, more than half of all possible cut scores (15), the misplacement will be equal to zero. In other words, the misplacement is minimal in the Level Characteristic Curves method. The mean misplacement (absolute value) is 0.82 which is lower than the mean misplacement for all of the other five methods. In other words, the Level Characteristic Curve method is better than the rest regarding the criterion of minimum distortion and would work well in case of

adequate judgments. Unfortunately, the judgments are not always adequate.
The Level Characteristic Curves method, however, just as the Basket procedure method, relies only on the judgment in setting cut scores without making corrections to take account of the empirical data and this is its main **disadvantage**.
The other **limitations** of the method are the narrow scope of application due to the requirement that the test it is applied to must be developed using IRT along with the need for more complex statistical software.
The main **advantage of the method** is the minimal deviation of the obtained cut scores from the expected scores. Other advantages are the clear interpretation of the results, as well as the possibility for additional analyses of the judgment by levels, and the graphical presentation of the data for setting cut scores.
Keeping in mind the advantages and disadvantages of this method it can be concluded that the most appropriate use of the Level Characteristic Curve method is as a secondary method, mainly in the analysis of the validity of the judgments and of the cut scores set through some other method.

3 Comparative analysis of the quality of the separate methods

3.1 Methods

3.1.1 Goal, main purposes, object, subject and hypotheses of the current research

One of the most significant conclusions from the many years of research into the methodology of setting cut scores is that the different methods lead to different cut scores (Jaeger, 1989; Mehrens, 1994; Bontempo et al, 1998 and others). This was confirmed while comparing the six methods in Chapter Two where even for the very same judges the cut scores were different for the different methods. For some of the methods the cut scores diverged to a large extent.

These differences are easy to explain. They emerge from the fact that the procedures in the different methods are also different and take into account different kinds of information. In the absence of a true cut score, it is difficult to make a decision regarding which of the different cut scores is the most adequate and well-supported. Existing empirical data and theoretical evidence and reasoning are indispensible for supporting the validity of a given method and its advantages compared to the other methods.

Accordingly, the **main goal** of this empirical study is to *investigate the validity of the six methods developed by the author for setting cut scores and based on predetermined criteria and a comparative analysis using these criteria to determine the most effective among the methods.*

The subject of the present research is the methods for setting cut scores, described in Chapter Two, namely:

- *The Basket procedure;*
- *The Compound Cumulative method;*
- *The Cumulative Cluster method;*
- *The ROC-curve method;*
- *The Item Mastery method;*
- *The Level Characteristic Curve method.*

The object of the present study is the psychometric characteristics of the aforementioned methods, in terms of their validity (internal and external).

As already mentioned in setting the research question (Chapter One – 1.1.), in the process of development of the aforementioned methods, several **hypotheses** were outlined, namely:

I. The method with the smallest standard error of the cut score is the **Compound Cumulative method**.

II The cut scores obtained by the **Basket procedure** will differ substantially from the ones obtained using the other five methods, and the direction of these differences will depend on the position of the respective cut score on the scale used for measuring the test results.

III The **Compound Cumulative** and **Cumulative Cluster** methods will produce cut scores closer to each other than the other methods.

The main purposes of the study are as follows:

A Development of a study design;
B Application of the six methods for setting cut scores for each particular component of the study design;
C Analysis of the internal validity of the obtained cut scores and testing hypothesis I;
D Comparative analysis of the cut scores, determining the degree of proximity between the separate methods and testing hypotheses II and III;
E Development of a set of criteria to determine the quality of the methods for setting cut scores which conform to the characteristics of the study design and the prevailing criteria for quality;
F Comparative analysis of the methods for setting cut scores in the light of the system of criteria and determining the most effective among them.

The successful realization of the aims described in A to F will be directly related to attaining the main goal of the study and to testing the hypotheses described above.

3.1.2 Study design

The study has been designed to control the factors that might affect the cut scores but which at the same time are not directly connected to the subject of the analysis.

To avoid the influence of the 'judge' factor, which in general is crucial, it is desirable to set the cut scores on the basis of the same judgments. In this case it is possible because the six methods are based on the same type of judgment.

Moreover, the specifics of the methods used in the study make it possible to conduct secondary analyses of data from projects in which only one of the described methods is used. This can be achieved by simply using the available empirical data and judgments to set the cut scores with each one of the other methods.

This, however, requires a decision to be made as to which of the available data to use in the secondary analysis, as over the last ten years a significant amount of empirical data has been obtained by the author: over 30 tests for assessing the level of language competence in 7 languages (English, German, French, Spanish, Swedish, Finnish and Russian), four reproductive language skills (reading comprehension, listening comprehension, grammar and vocabulary), and the overall number of judgments that have been taken into account is over 400.

In the end, the choice concerning the particular empirical data to be included in this study was made on the basis of several criteria:

- High reliability of the test and good psychometric characteristics of the items that it contains.
  → This requirement comes from the fact that the validity of the cut scores is highly dependent on the validity of the test itself. If the test is of doubtful quality, it cannot be expected that the cut scores will be adequate and valid. Moreover, the classification accuracy in dividing the test score into several groups is directly related to the reliability of the test.

- Tests which have a good fit between the empirical data and the chosen IRT model.
  → Since two of the methods are IRT-based, the empirical data has to be such that it allows the application of some of the IRT models. In turn, a high degree of consistency between the model and the data is necessary because any violation of the conditions for using a given model casts into doubt all the subsequent conclusions and interpretations, including those concerning the cut scores obtained by using the model.
- A maximum variety of language abilities being measured.
  → It is well-known from practice that the quality of achievement tests is predetermined to a large extent by the complexity of the construct being measured (cognitive ability or competence). For instance, in foreign language testing it is relatively easier to develop a good test for measuring vocabulary or grammar knowledge than a test for measuring reading or listening comprehension. It is also logical to suppose that the judges, most of whom are also test item writers, will show a different degree of consistency with the empirical data depending on what exactly is the language ability that is measured. That is why, if the purpose is to explore the effectiveness of the methods, regardless of the ability measured by the test, it is desirable to apply the methods to tests measuring different abilities.
- Variation in the number of the judges participating in the judgment task.
  → The accuracy of the judgment, in the sense of a minimal error of the mean of the judgments of all judges, is one of the main criteria for the quality of a given method and the cut scores obtained by it. One of the main factors on which the standard error depends is the number of participating judges. That is why it is desirable that the empirical data is selected in a way that ensures variation in the number of judges that took part in setting cut scores.
- Variation in the quality of the judgments
  → The cut scores, regardless of the method applied, depend to a large extent on the quality of the judgments – whether and to what extent there is agreement between the different judges on the one hand and the judgments and the empirical data on the other. In order to explore the influence of the quality of the judgments on the cut scores derived by different methods, judgments of varying quality are required.

It is obvious that it is necessary to include more than one test in the present study in order to ensure the desired variety concerning the measured skills and the number of judges that took part. Taking into account these main considerations, the final design for the present research includes **three tests** and the related **judgments**. Their description is presented in the next two sections.

3.1.2.1 Instruments

In analysing the results of the three tests the same probability model – the one-parameter logistic model – OPLM – was used (Verhelst et al, 1995). This model unites the attractive characteristics of the Rasch model with the higher flexibility and applicability of the two-parameter logistic model. The main difference between the two models is that the one-parameter logistic model does not require the test items to have the same discrimination power, as is the case in the Rasch model, which is hard to achieve in reality. One of the

consequences of this difference is that, in the case of the one-parameter model, the total test score is based on test items with different weights which are proportional to their discrimination index. In other words, the raw test score and the Z-scale – used for expressing the results when using OPLM for a mutual correspondence that allows comparability – have to be treated as different measurement scales.

The first test (T1) is a test for listening comprehension in Finnish and is part of the international European project for Internet-based foreign language testing – DIALANG. The pilot study, with the judgment data and a preliminary analysis of the data, was conducted in the period 1998-1999.

The final version of the test consists of 50 test items, 36 multiple-choice items and 14 constructed response items.

The initial test characteristics were determined using a sample of 429 examinees and an incomplete linked test design of four blocks. In developing the adaptive sub-tests and the analysis of the cut scores a simulation model with 900 examinees was used.

The psychometric characteristics of the test are presented in Table 11 (column "Test 1"). As the table shows, the test demonstrates very high reliability (0.96), regardless of the fact that it contains items with a quite low discrimination index (+0.11). The fact that such items remained in the final version of the test has its explanation: the test is intended to measure the language competence in the entire interval from level A1 to C2 using adaptive subtests. That is why it contains items with a high variability in their difficulty. For instance, the item with a discrimination index of +0.11 is one of the easiest items and was answered correctly by 98% of the examinees. However, it has a positive discrimination index, a monotonously increasing Level Characteristic Curve and shows good fit to the theoretical model that was used.

Table 11 Psychometric characteristics of the tests

Indicators		**Test 1**	**Test 2**	**Test 3**
Language		Finnish	English	Swedish
Ability		Listening	Reading	grammar
Number of examinees		429 (900)	2622 (277)	15370
Number of items		50	52	39
Difficulty	Minimum	23%	27%	13%
	Mean	67%	64%	68%
	Maximum	98%	90%	96%
Discrimination index	Minimum	0.11	0.19	0.24
	Mean	0.40	0.48	0.46
	Maximum	0.65	0.72	0.64
Raw test score	Maximum	50	52	39
	Mean	31.98	33.11	26.63
	Standard deviation (*SD*)	13.12	11.00	7.45
Reliability (α)		**0.96**	**0.93**	**0.90**
Standard error (*SEM*)		**2.62**	**2.87**	**2.36**
Test score (Z-scale)	Maximum	2.853	2.833	2.291
	Mean	0.415	0.390	0.353
	Standard deviation (*SD*)	0.796	0.536	0.499
Fit between the model and the data	Items (*p*)	$p > 0.012$	$p > 0.014$	$p > 0.009$
	Test (*p*)	$p = 0.111$	$p = 0.146$	$p = 0.066$

The second test (T2) is one of the pilot tests for grade 11 used in the Norwegian project for external assessment of language competence (reading comprehension) in English in secondary school (BITE-IT: Bergen Interactive Testing of English).
It contains 52 items and is completely computerized (including the item scoring). The item format is compatible with IT capacities and is close to the formats used in the items in the illustrative example (section 2.2.1). The empirical data for this test were also obtained using an incomplete linked test design containing 10 blocks and a total sample of 2622 examinees. Each test item was answered by at least 513 examinees. The initial test characteristics were determined using the whole sample, but the analysis of the cut scores was based only on a sub-sample of 277 examinees who answered all the 52 items in the test.
The psychometric characteristics of the test are presented in Table 11 (column "Test 2") and it shows that the test has a high reliability (0.93) and good fit (p = 0.146) with the theoretical model used (OPLM).
The third test (T3) is a sub-test for the assessment of language competence (grammar structures) in Swedish that was used for the matriculation examination (of Finnish-speaking students) in Finland in 2004. The sub-test contains 39 multiple-choice items in total. The psychometric characteristics of this instrument were determined using the total population of students that took the examination (n = 15 370).
Due to the size of the sample, however, the results for fit between the theoretical model and the data (Table 11, the last two sections of column T3) are based on the random sub-sample of 500 examinees. The reason for this sub-sampling is that all the tests of statistical significance are highly dependent on the sample size and with $n > 1000$ even minimal differences are statistically significant.
It should be noted that for this test, because of security concerns, namely the need to ensure that item content was not leaked prior to the administration of this examination, the test items were not pre-tested. Despite this fact, however, the test shows a high quality (α = 0.90) keeping in mind its relatively short length (39 items). Moreover, even the minimum discrimination index (+0.24) of all items is acceptable as it is for one of the most difficult items that was answered correctly by only 31% of the examinees.

3.1.2.2 Judgments

The judgments for setting cut scores for the three tests included in the study were obtained using the same formulation of the judgment task, namely:
Which is the minimum level of competence at which a given examinee has to be in order to answer this test item correctly?
Each one of the judges gave his or her judgment for each one of the test items and was free to use the entire scale of the language competence levels (A1, A2, B1, B2, C1 and C2) as defined by the CEFR. Depending on the measured language ability, different sub-scales were used. Due to the lack of a specific scale for the assessment of grammatical knowledge in the CEFR, the corresponding scale developed in Finland was used in the third test.
This scale is linked to the CEFR.
All of the judges that participated in the judgment process were experts in the respective foreign language with over two years of experience both in teaching and in test item

construction and analysis.
The preliminary instruction of the judges in all three cases was conducted in half a working day and included introduction to the CEFR and the corresponding assessment scale as well as judgment of sample test items in order to increase consistency with the empirical data among the judges
The number of judges that took part in setting the cut scores for the first test was seven. This number, although exceeding the absolute minimum of five judges (Livingston & Zieky, 1982, p. 16) and acceptable in forensic practice in the United States (Biddle, 1993), is relatively low. The reason for this low number is that the test is in Finnish and the number of experts with a minimum two years of professional experience in instruction in Finnish and testing is relatively low.
In the judgment for the second test, 10 judges took part which is in accordance with the recommendation given in the pilot version of the Manual for Relating Language Examinations to the CEFR (Council of Europe, 2003, p. 94).
The largest number of judges (13) participated in setting the cut scores for the third test and it would be expected that the standard error of the mean for the judgments for this test would be lower than in the other two tests.
The standard error of the mean of the judgments, however, depends not only on the number of judges, but also on their mutual internal consistency as well as on their consistency with the empirical data.
Table 12 presents the values for three of the most frequently used indicators of internal consistency – Cronbach's alpha (α), the intraclass correlation (ICC) and Kendal's coefficient of concordance (Kendall's W). In contrast to the other two indicators, Cronbach's α depends to a large extent on the number of the judges and as such it would be expected to be highest for the third test. However, the results in Table 12 do not meet this expectation. For all three indicators, the highest values are for the judgments made with the lowest number of judges (test 1) and the lowest for the third test, which has the most judges. Based on these results, it can be concluded that the judgments for the first and the second tests are of acceptable quality regarding their internal consistency. Concerning the judgments for the third test, the discrepancies between the judges are so large that we should talk about internal inconsistency rather than consistency. There are many possible reasons for such inconsistency which would be of interest to investigate, but in this case we are more concerned with investigating what happens when we apply the different standard-setting methods in this study to data showing such a low degree of internal consistency.

Table 12 Judgments

Indicators			**Test 1**	**Test 2**	**Test 3**
Number of judges			7	10	13
Number of items			50	52	39
Cut score **X** between levels:			A2/B1	B1/B2	B1/B2
Cut score **Y** between levels:			B2/C1	B2/C1	B2/C1
Internal consistency		α	0.95	0.94	0.85
		ICC	0.69	0.61	0.30
		W	0.78	0.63	0.36
Consistency with the empirical data (Z-scale)	ρ – Spearman's coefficient	Minimum	+0.34	+0.51	-0.02
		Mean	**+0.56**	**+0.56**	**+0.36**
		Maximum	+0.84	+0.61	+0.59
	τ_b – Kendal's coefficient	Minimum	+0.38	+0.28	-0.03
		Mean	**+0.42**	**+0.46**	**+0.28**
		Maximum	+0.46	+0.72	+0.48
	MPI – misplacement index	Minimum	0.73	0.68	0.48
		Mean	**0.76**	**0.79**	**0.68**
		Maximum	0.79	0.96	0.82

The degree of consistency with empirical difficulty (Z-scale) also varies too much – so much that in Test 3 there is one judge whose judgment is inconsistent with the item difficulty in the majority of cases, which leads to a negative (although not statistically significant) correlation: ρ = -0.02; τ_b = -0.03 and MPI = 0.48. At the same time, in Test 2, there is one judge whose judgment is in almost absolute consistency (96%) with the empirical item difficulty.

The degree of consistency of the judgments with the empirical item difficulty – across all three measures of consistency – is lowest (although acceptable) for the judgments for the third test.

The calculation of the different indices of consistency with the empirical data allows an analysis of the relationships between the different coefficients. The extremely high correlation (≥ 0.98) between the three indices (ρ, τ_b and MPI) confirms that they measure the same characteristic. In this, as could be expected, the relationship between the misplacement index (MPI) and Kendal's coefficient (τ_b) is relatively stronger ($r_{(\tau,MPI)}$ = 0.99) than between the MPI and Spearman's coefficient ($r_{(\rho,MPI)}$ = 0.98).

Although, for some of the tests in the preliminary analysis, more than two cut scores were set (for Test 1 and Test 3), for the needs of this study only two cut scores per test will be set. The corresponding levels of language competence are presented in Table 12, and the two cut scores will be designated by ***X*** (the first) and ***Y*** (the second).

3.1.2.3 Resampling

Since in setting the cut scores, relatively small samples of judges (n < 20) are used, one of the main problems is to find whether in any replication of the procedure with another sample of judges the results will be the same or at least close.

The comparative analysis of the results from the application of different methods is also

limited by the relatively small number of observations and the large number of factors on which they are highly dependent.
For instance, the total number of cut scores that are set for each test using the six methods is 20. This is so because two cut scores were set (X and Y) with each method, and the first four methods were applied both to the raw test score and the Z-scale, and the last two methods were used for setting the cut scores using only the Z-scale. This means that, with three tests, the number of the different cut scores that will be subject to a comparative analysis is 60, which is too small a sample size, especially as these 60 cut scores will be affected by several factors such as the test, the method, the sequential number of the cut scores and the type of scale.
To overcome these two problems, modern statistical methods offer several possible approaches that are known under the common term **resampling**. Resampling is viewed as a new and promising alternative to the statistical tests of significance (Yu, 2003; Rodgers, 1999) and is suitable especially in cases of small samples, as in the setting of cut scores. Unfortunately, it is still not widespread in this area and its only known application until now has been in examinee-centered methods (Muijtjens, A. et al., 2003).
The *'jackknife'* procedure (Ang, 1998; White, 2000) was used as the method of resampling in this study, as follows:
From each initial sample of *n* judges, sub-samples are drawn withoutreplacement.
The number of elements in the new sub-samples will number one less than the original sample. The number of the different combinations of *n* elements from class *n*-1 is equal to *n* and hence the number of these sub-samples will be equal to the number of cases in the initial sample.
Since the original data consisted of three samples of judges whose number is 7, 10 and 13 respectively, using the *'jackknife'* procedure 30 new sub-samples were drawn and for each one of these samples the number of the respective cut scores was 20. This led to a total of 600 cut scores which is a sufficiently large sample of observations. Based on this sample well-grounded statistical conclusions can be drawn.
The *'jackknife'* resampling procedure also allows an additional evaluation of the parameters of interest (in this case cut scores) and to judge their stability (*replicability*) in any repeated application with the same number of judges (Thompson, 1994; Gillaspy, 1996; Kier, 1997; White, 2000).

3.2 Empirical results

3.2.1 Cut scores

The two cut scores for each judge depending on the corresponding assessment scale (raw test score or Z-scale) and the method that was used are presented in Appendices 6A and 6B. Figure 28 graphically presents part of these data – the cut scores set by the different judges using the Basket procedure. These scores correspond directly to the frequency distribution of the items assigned to the separate levels (≤ X; (X; Y]; >Y), by the different judges.
The data in Appendices 6A and 6B show clearly that there was a strong variation in the judges' opinion, and there were cases in which the first cut score (X) for one judge was

higher than the second cut score (Y) for another judge (e.g. E6 and E7 in Test 3). Strongly deviating results (*outliers*), however, were evident only in Test 2 for E4.
There were also some tendencies which are quite notable, such as the higher heterogeneity of the judges in setting the lower cut score and the relatively more homogeneous judgments for the first test. The last tendency is easy to explain since the internal consistency for the judges is highest for the first test and lowest for the third (Table 12).

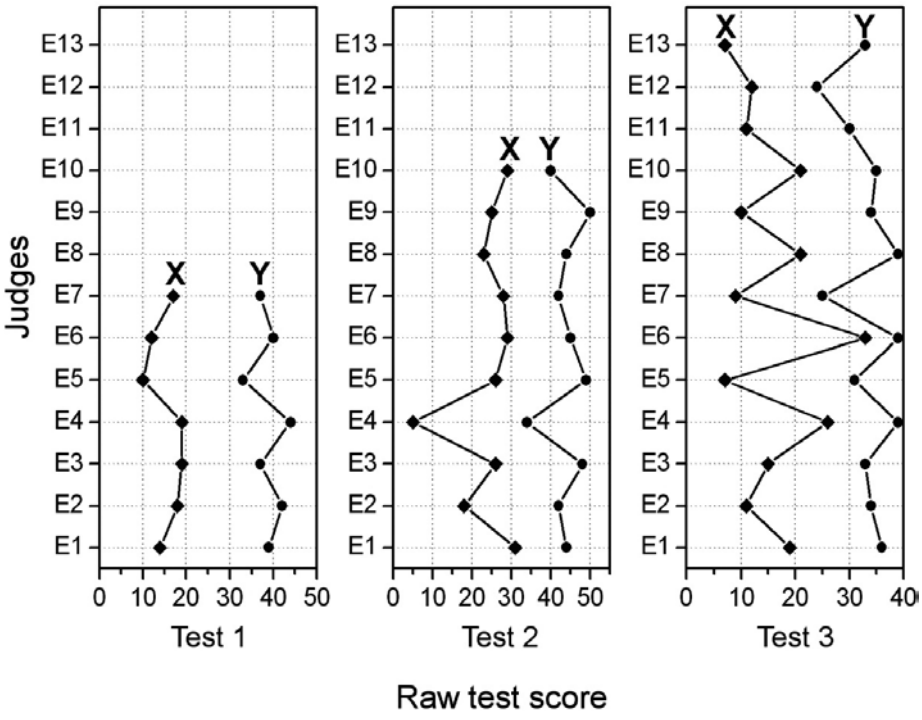


Figure 28 Judgments: Basket procedure (raw test score)

There are several possible reasons for this variation in the obtained judgments. The main reason is the subjective nature of the judgment and the differences in the values and the judgment styles of the different judges. A second reason, which is of no less importance, is related to the training of the judges, the main purpose of which is to maximize a common interpretation of the judgment criteria and to make the judgments homogeneous. The fact that even after such training there was a difference of 26 points between the cut scores of the different judges (Test 3 – judges № 6, 5 and 13), with a maximum test score of 39 points, simply means that the training was not effective enough.
Unfortunately the high heterogeneity of the judgments is not an isolated phenomenon or restricted to these three particular cases. A meta-analysis of 90 comparative studies employing two different methods for setting cut scores showed that *"the variability within a standard setting method is at least as large as any difference between standard setting methods"* (Bontempo et al, 1998, p. 10).

With such high variability of the judgments and relatively small numbers of judges (usually lower than 15), there is good cause to argue for using the median of the judgments when setting the final cut score. The reason for this is that this measure of central tendency does not depend directly on the results that deviate strongly from the mean, which in relatively small samples can itself be strongly influenced by potential extreme values.

Using any of the measures of central tendency has both advantages and disadvantages (Livingston & Zieky, 1982, pp. 21-22). Moreover, there is still a lack of empirical studies concerning the effect of using particular measures of central tendency on the obtained cut scores and the quality of the classification decisions (Karantonis & Sireci, 2006, p. 10).

The main argument against using the median is that its standard error is about 25% higher than the standard error of the mean (Jaeger, 1991, p. 6). This argument carries considerable weight because, as will be shown later, the size of the standard error of the cut score is one of the main criteria for the quality of a given method.

Another argument against using the median of the cut scores of the different judges as the final cut score is that this would lead to a lot of information being lost. This happens because the median is equal to the cut score that is in the middle of the interval of all cut scores, hence the judgments of all judges, except the middle one, do not play a role in the final cut score, i.e. they are excluded in its setting. In practice this approach means that, in cases in which a considerable number of judges participate, their judgments are not taken into account in setting the final cut score.

This would effectively exclude the judgments of a considerable number of judges, and according to Jaeger (1988, p. 29) not including data from participants whose judgments are in fact relevant to setting the cut score on the test in question runs counter to the aim of using the judgments of a panel of experts to derive the cut score. This could lead to doubts concerning the adequacy of the obtained cut scores (Norcini, 2003, p. 467).

That is why in all the methods that are the subject of this research the final cut scores will be equal to the mean of the cut scores of all judges that participated in the particular judgments. These cut scores ($\boldsymbol{C}$), as well as their standard error ($\boldsymbol{SE_c}$) and standard deviation ($\boldsymbol{SD_c}$) are presented in Appendix 6C, and the graphics in Figure 29 graphically present their mutual location on the scale used for presenting the test results.

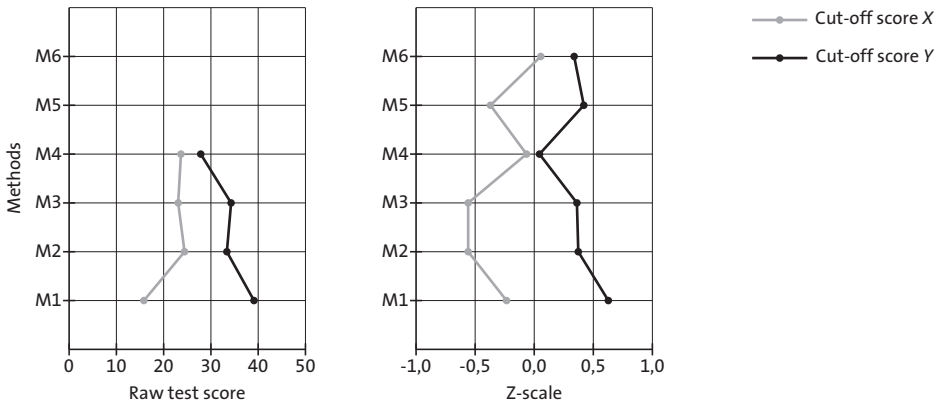


Figure 29a Final cut scores for Test 1

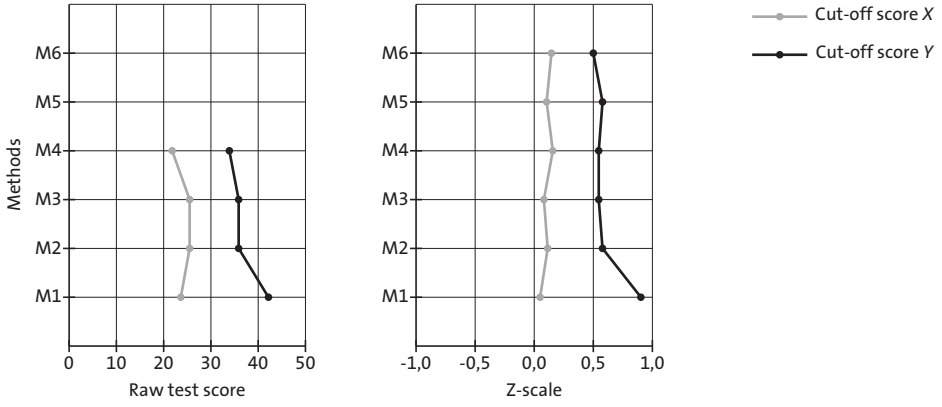


Figure 29b Final cut scores for Test 2

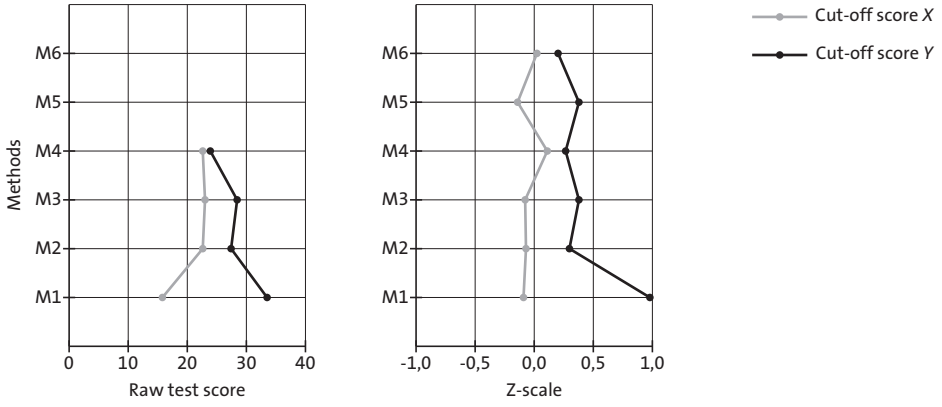


Figure 29c Final cut scores for Test 3

The comparative analysis of these cut scores shows that they vary greatly depending both on the method used for their setting and on other factors – test, sequence of the cut score (first – X or second – Y) and the scale (raw test score or Z-scale). At the same time there are some common tendencies, namely:

- The first cut score (*X*) in the Basket procedure (M1) is usually lower than in the other methods.
- Exactly the opposite tendency is observed for the second cut score (*Y*), which is the highest in the Basket procedure (M2).
- There is a tendency for the two cut scores (*X* and *Y*) to become closer to each other in the ROC-curve method (M4) and the Level Characteristic Curves method (M6). This tendency is clearer in M4 as the difference between the two cut scores for the raw test scores (X_{M4} = 22.23 and Y_{M4} = 23.40) for the third test is lower than the standard error of the measurement for this test (*SEM* = 2.36).
- The cut scores obtained through the Cumulative Compound method (M2), the Cumulative Cluster method (M3) and the Item Mastery method (M5) are closest to each other.

The same tendencies were observed also for the cut scores obtained by resampling. These are presented in the lower part of the table in Appendix 6C.

3.2.2 Replicability and precision of the cut scores

3.2.2.1 Degree of matching in resampling

It is obvious that the cut scores in the two halves (upper and lower) of the table in Appendix 6C are very close to each other. One of the criteria for stability (Gillaspy, 1996; Ang, 1998), which is usually used in resampling, is the value of the *t*-statistics for a given parameter (in this case – the mean cut score), which should be higher than the critical value of *t* with *n*-1 degrees of freedom and the corresponding level of confidence probability (in this case 95%). The values of *t* are located in column t_c, and the ones that are lower than t_{cr} are bolded and in shaded cells. As can be seen, from a total of 60 cut scores, only 7 are not stable enough and all of them concern the Z-scale in cases where the mean value is close to zero. In other words, the lower stability is mainly due to the specific characteristics of the cut scores in these particular cases.

This conclusion is supported by the second criterion for replicability (Gillaspy, 1996; Ang, 1998). According to this criterion, the parameter of interest is stable (replicable) if the judgments obtained using the original sample are within the 95% confidence interval of the judgments obtained by resampling. This condition is satisfied for each one of all the 60 cut scores. Moreover, as Figure 30 shows, the pairs of cut scores (original and resampled) are located very close to (and over) the diagonal for which the equation $y = x$ is true.

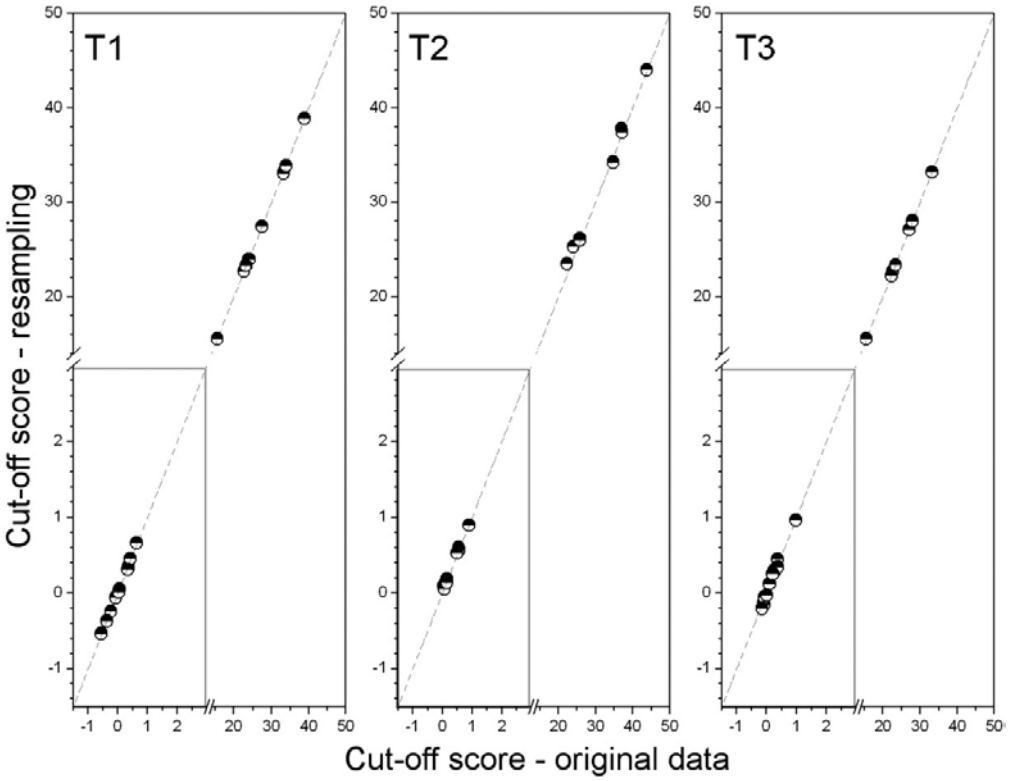


Figure 30 Replicability of the cut scores

Some, although insignificant, displacement from the diagonal is observed mainly for the second test and this can be explained by the presence of strongly deviating results for one of the judges (E4).
Based on these convincing results it can be concluded that despite the small samples of judges and the heterogeneity of the judgments, the final cut scores are replicable. That is why it can be argued that in a potential replication of the judgments the obtained results will be close to the ones in this study.

3.2.2.2 Standard error of the cut scores

Standard error is one of the classical indices indicative of the replicability of the obtained results. The higher the standard error, the higher is the probability of significant differences in the obtained cut scores in a replication of the judgment with different judges.
Moreover, the classification accuracy of the test depends on two main factors – the standard error of measurement (*SEM*) and the standard error of the cut score (SE_c). This is because the classification category to which a given examinee will be assigned depends not only on the degree of precision of the measurement of his or her competence, but also on the degree of precision in setting the cut scores for the test. An indicator for the classification accuracy of the test is the so called *'total standard error'* (SE_{tot}), which is equal

to the square root of the sum of squares of *SEM* and SE_c (Jaeger, 1991, p. 6). The standard error of measurement for a given test, however, is a fixed value, although it can take different values for the different test scores. That is why the classification accuracy will depend mainly on the standard error of the cut scores whose variation depends on the method used for obtaining it.

This is the reason why the standard error of the cut score is accepted as one of the main indicators for the validity of the corresponding cut score and for the quality of the corresponding method (van der Linden, 1994, p. 10; Kane, 1994, pp. 445-446; Hambleton & Pitoniak, 2006, p. 460; Reckase, 2006, p. 6; Cizek & Bunch, 2007, pp. 60-61 and others).

What is an acceptable value of the standard error of the cut score is a question that does not have an exact answer. Nevertheless, there are several tentative criteria. They are based on the evaluation of the total standard error SE_{tot} in relation to the corresponding standard error of measurement. For instance, if $SE_c \leq \frac{1}{3}SEM$, then SE_{tot} will increase by no more than 5%, compared to when the cut score is derived without any error ($SE_c = 0$) and $SE_{tot} = SEM$. If, however, $SE_c > SEM$, then the increase in the SE_{tot} will be over 41%, hence the error of the cut score will have a relatively large impact on the classification accuracy.

To reduce this effect to a minimum, Jaeger (Jaeger, 1991, p. 6) recommends that the standard error of the cut scores be no higher than a quarter of the standard error of measurement ($SE_c \leq \frac{1}{4}SEM$). This would lead to an increase of no more than 3% in SE_{tot}. This criterion is extremely strict and is very hard to meet in reality if the number of judges is lower than 20.

That is why Cohen, Kane and Crooks (Cohen et al, 1999, p. 364) suggest a more liberal criterion – $SE_c \leq \frac{1}{2}SEM$. The percentage increase of SE_{tot} in this case, compared to the hypothetical case of $SE_c = 0$ and $SE_{tot} = SEM$, would be no more than 12%.

A compromise between these two criteria is the first criterion mentioned above of $SE_c \leq \frac{1}{3}SEM$. With this criterion the increase is no more than 5%, which means a minimum effect of SE_c on the classification accuracy. Meeting this criterion is possible with a limited number of judges (≤ 15), as the results of this study will show.

The standard errors for each cut score are presented in Appendix 6C, both for the original data and for the resampling. These standard errors, however, are expressed in different measurement scales (raw test-score and Z-scale) and do not allow a comparison between the standard errors of the cut scores for the different tests and different scales. To provide the opportunity for such comparison, a new variable (*SE/SEM*) was introduced. It is equal to the ratio of a given standard error of the cut score (SE_c) and the corresponding standard error of measurement (*SEM*). Since, however, the standard error of measurement is different for the different points of the measurement scale, in calculating the values for the new variable (*SE/SEM*), instead of the mean standard error *SEM*, the conditional standard error for this value of the measurement scale that matches the corresponding cut score is used. For calculating the conditional standard error Keats's modification of Lord's binomial approach was used (Feldt et al, 1985, pp. 353-354).

The new variable (*SE/SEM*) not only gives the opportunity for a comparison of the standard errors for the cut scores expressed in different scales and for different tests, but also allows for a quick test of whether the corresponding standard error is congruent with the criteria for quality that were set earlier. For example, if $SE/SEM \leq 0.25$, then the standard error of

the corresponding cut score meets Jaeger's criterion. In the same manner, if *SE/SEM* ≤ 0.5, then the standard error of the corresponding cut score will meet the criterion of Cohen, Kane and Crooks.
Table 13 shows the mean values of this new variable broken down by the method for setting cut scores. The analysis was conducted both for the original 60 cut scores and for the 600 obtained by the resampling. As the table shows, the standard errors of the cut scores are minimal for the Cumulative Compound method (M2) which **supports the first hypothesis** of the current research.

Table 13 Ratio between the standard error of the cut scores and the standard error of the measurement

Data	Method	Number of cases	SE/SEM		Citerion				Mann-Whitney test Z (M_2 & M_i)	Statistical significance ($p > 0.05$)
			Mean	*SD*	≤ 0.25	≤ 0.30	≤ 0.50	≤ 1.00		
Original	M_1	12	0.71	0.20	no	no	no	**yes**	3.61	**0.000**
	M_2	12	**0.32**	0.13	no	**yes**	**yes**	**yes**	-	-
	M_3	12	0.37	0.15	no	no	**yes**	**yes**	0.43	0.434
	M_4	12	0.67	0.24	no	no	no	**yes**	3.24	**0.001**
	M_5	6	0.49	0.12	no	no	**yes**	**yes**	2.31	**0.021**
	M_6	6	0.38	0.19	no	no	**yes**	**yes**	0.70	0.481
Resampling	M_1	120	0.74	0.20	no	no	no	**yes**	12.19	**0.000**
	M_2	120	**0.32**	0.14	no	**yes**	**yes**	**yes**	-	-
	M_3	120	0.35	0.16	no	no	**yes**	**yes**	1.46	0.144
	M_4	120	0.66	0.24	no	no	no	**yes**	10.17	**0.000**
	M_5	60	0.56	0.14	no	no	no	**yes**	8.51	**0.000**
	M_6	60	0.43	0.18	no	no	**yes**	**yes**	4.09	**0.000**

Second in terms of the degree of accuracy is the Cumulative Cluster method (M3), as the difference between the mean values of *SE/SEM* for it and M2 is not statistically significant (the last two columns of Table 13). The same applies to M2 and M6.
Statistically significant, however, are the differences between the Cumulative Compound (M2) and the other three methods (Basket procedure – M1; ROC-curve method – M4; Item Mastery method – M5).
These conclusions apply both to the original data and for the resampling.
From the point of view of the standard error, it is obvious that the lowest quality is shown by the Basket procedure (M1) and the ROC-curve method (M4). There is no statistically significant difference between them ($p = 0.686$), but they display a statistically significant difference when compared to the other four methods ($p < 0.05$). This is confirmed by the confidence intervals for the mean values of *SE/SEM* for the cut scores from the resampling (Figure 31).

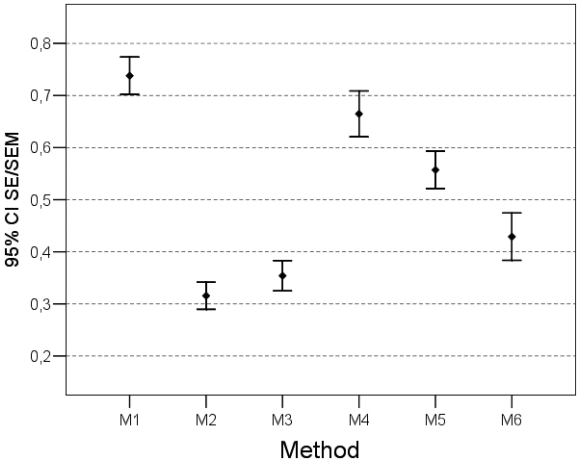


Figure 31 Confidence intervals for SE/SEM (resampling)

This figure is also a good illustration of the quality of the separate methods concerning the defined criteria for the precision of the cut scores. Obviously the mean standard errors for none of the methods meet Jaeger's criterion ($SE_c \leq \frac{1}{4}SEM$). This is mostly due to the small samples of judges both for the original data ($n_1 = 7$; $n_2 = 10$ and $n_3 = 13$) and for the data from the resampling where the sample sizes have numbers one less than the original sample sizes.

The criterion $SE_c \leq \frac{1}{3}SEM$ is met only by the mean standard error of the Cumulative Compound method (M2). Concerning the criterion of Cohen, Kane and Crooks, ($SE_c \leq \frac{1}{2}SEM$), only the mean standard error of the Basket procedure (M1) and the ROC-curve method (M4) do not meet it for the original data, while for resampling the Item Mastery method (M5) also fails to meet this criterion.

The standard errors of the cut off scores for the Cumulative Compound (M2) and Cumulative Cluster (M3) support (although not directly) the third hypothesis of the present study. The final test of this hypothesis will be made later.

The results of table 13 and Figure 31 are based only on the mean values of the variable *SE/SEM*, which could give an imprecise picture, especially with the original data, where the number of the cut scores for each method is limited. That is why Figure 32 presents the values of *SE/SEM* for each of the 60 original cut scores that are analyzed in the present study.

This figure shows that the values of *SE/SEM* for almost half (5 out of 12) of the cut scores obtained through the Compound Cumulative method (M2) are lower than 0.25 *SEM*, hence, they meet even the extremely strict criterion of Jaeger. In addition, the Cumulative Compound (M2) and the Cumulative Cluster method (M3) are the two methods in which the number of the cut scores that do *not* meet at least the Cohen, Kane and Crooks criterion is the lowest (17%). For the rest of the methods this percentage is as follows: 33% for the Level Characteristic curve (M6); 50% for the Item Mastery method (M5); 75% for the ROC-curve method (M4) and 92% for the Basket procedure (M5).

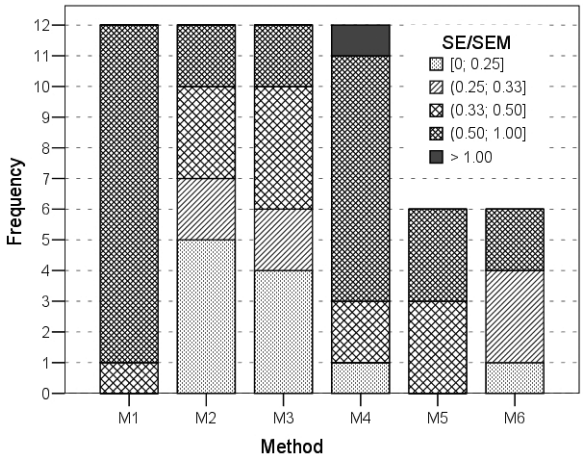


Figure 32 Standard errors of the cut scores and criteria for quality

The main conclusion is that in most cases the standard error meets the criteria set by Cohen, Kane and Crooks for $SE_c \leq \frac{1}{2}SEM$ for three of the methods, namely:

- the Cumulative Compound method – M2 (83 % of the original cut scores and 83% of the resamplings);
- the Cumulative Cluster method – M3 (83 % of the original cut scores and 77% of the resamplings);
- the Level Characteristic Curve method– M6 (67 % of the original cut scores and 63% of the resamplings).

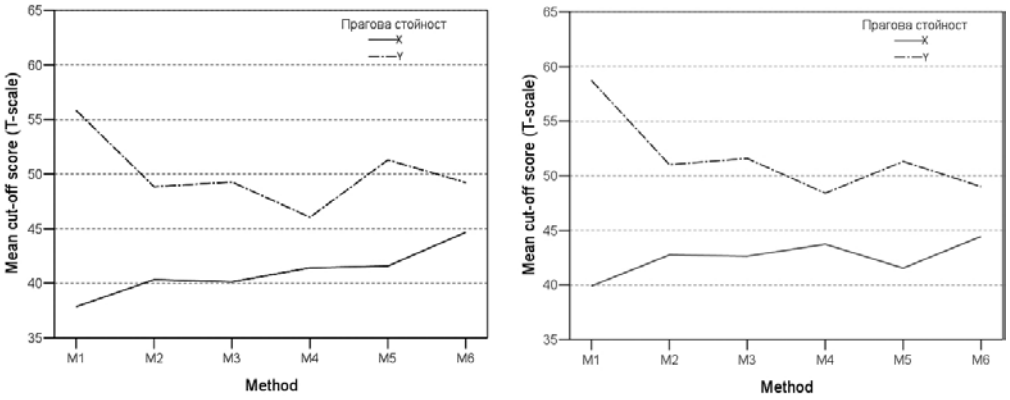
Another conclusion that can be drawn from Figure 32 is that the ROC-curve method (M4) has the highest variation of the standard error of the cut scores. This accordingly also means the highest unpredictability in its application because its standard error can be both extremely small (SE_c = 0.01 and *SE/SEM* = 0.1 for T1; cut score *X* in *Z*-scale) and extremely large (SE_c = 0.15 and *SE/SEM* = 1.1 for T2; cut score *Y* in *Z*-scale). This conclusion is confirmed also by the standard deviation of SE/SEM (Table 13), which is highest for this method. In this context, in terms of the standard error, the most homogeneous are the Cumulative Compound method (M2), the Cumulative Cluster method (M3) and the Item Mastery method (M5).

3.2.3 Commensurability of the cut scores

3.2.3.1 Common tendencies

In presenting the final cut scores (3.2.1. – Figure 29), a few common tendencies were noted and they need additional confirmation. The difficulty in doing this comes from the fact that the separate cut scores are expressed on different measurement scales (raw test score or Z-scale) even for the same test. That is why, to provide comparability of the separate cut scores, they were all transformed into a T-scale using the formula: $\mathbf{T = 50+10.(C_i - M_i)/SD_i}$, where $\boldsymbol{C_i}$ is the corresponding cut score, and $\boldsymbol{M_i}$ and $\boldsymbol{SD_i}$ are the mean and the standard

deviation for the scale in which C_i is expressed.
This transformation gives the opportunity for comparison and obtaining the mean cut scores from the ones derived by different tests, methods and measurement scales. Such a comparison is provided by Figure 33 where the obtained mean cut scores for the different methods expressed on a T-scale are presented.



(a) Original data (b) Resampling

Figure 33 Mean cut scores depending on the method

As the graphs show, all previously noted tendencies are present again.
The cut scores X obtained with the Basket procedure (M1) are lower than the cut scores obtained with the other methods, and the cut scores Y obtained with the same method are higher than the other methods. This is true both for the original cut scores and the resampling and **supports the second hypothesis**.
This also supports the view expressed by Reckase, which is shared by other leading researchers in this area as well (Reckase, 1998; Hambleton & Pitoniak, 2006; Cizek & Bunch, 2007), according to which Angoff's alternative method and its modifications lead to a distortion of judgments – underestimation of the test scores lower than the mean and overestimation of the higher ones.
Concerning the significance of the differences between the cut scores obtained with the Basket procedure (M1) and the other methods, the lack of overlap in the confidence intervals for the mean cut scores obtained through the resampling (Figure 34) proves clearly that these differences are statistically significant at a 95% confidence level.

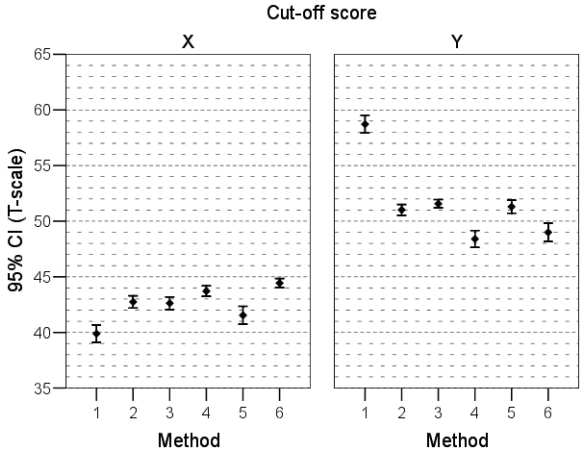


Figure 34 Confidence intervals for the mean cut scores (resampling)

The graphs in Figures 33 and 34 **support the third hypothesis of the present study** concerning the maximum closeness and statistically insignificant differences between the cut scores obtained by the Cumulative Compound method (M2) and the Cumulative Cluster methods (M3).
This closeness between the cut scores derived using the two different methods (M2 and M3) can be easily explained taking into account their common principles – repeated classification of the test items by levels of competence in correspondence with the mean item difficulty by levels according to the judgments.
The noted tendency of closeness of the two cut scores (X and Y) in the ROC-curve (M4) and the Level Characteristic Curve (M6) methods is also confirmed by the results shown in Figures 33 and 34. This tendency carries serious risks that cast the quality of both methods into doubt. The reason is that with a small difference between the two cut scores (compared with the standard error of the measurement – *SEM*) the decision at what level of competence a given examinee will be assigned will depend not so much on his or her real level of competence but mostly on the precision of the measurement instrument.
That is why the significance of the differences between two consecutive cut scores (X and Y) for one and the same method needs additional testing and this will be done in the next section.

3.2.3.2 Significance of the differences between cut scores X and Y

The difference between two values of the test score for the same test is statistically significant with a 95% confidence level if its absolute value is higher than 2.77*SEM*, where *SEM* is the standard error of measuement (Harvill, 1991, p. 38). In other words, the difference between the cut scores $\boldsymbol{C_x}$ and $\boldsymbol{C_y}$ will be statistically significant if the ratio $|\boldsymbol{C_x} - \boldsymbol{C_y}| / \boldsymbol{SEM} > 2.77$.
Figure 35 shows the mean values of this ratio for the different methods. As can be seen for the ROC-curve (M4) and the Level Characteristic Curve method (M5), this difference is not statistically different.

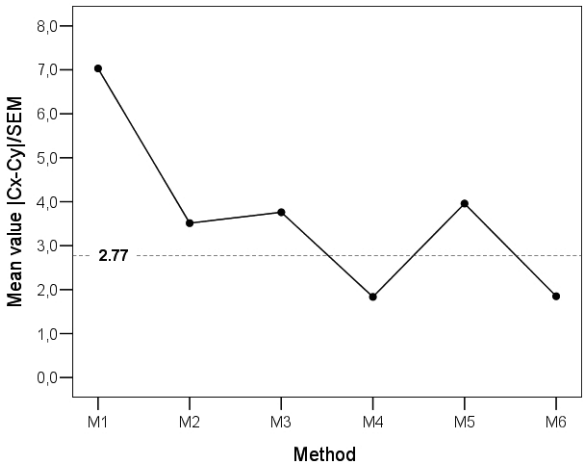


Figure 35 Significance of the differences between cut scores X and Y

This applies not only to the mean difference, but also to the majority of the various cases of comparison. For example, from a total of 6 differences for the ROC-curve method M4, (two for each test depending on the scale of measurement), four (75%) are lower than the critical value (2.77*SEM*). For the Level Characteristic Curve method, none of these differences (3 in total – one for each test) is statistically different.
This result leads to a very important conclusion: the ROC-curve method (M4) and the Level Characteristic method are not appropriate for setting more than one cut score for a given test, because the difference between two consecutive cut scores is usually not significant.

3.2.3.3 Significance of the differences between the different methods

One of the conclusions of the analysis of the common tendencies (section 3.2.3.1) was that there were no statistically significant differences between the cut scores obtained by the Cumulative Compound (M2) and Cumulative Cluster (M3) methods. This conclusion, however, was drawn using resampling and a comparison between the mean cut scores for the different methods expressed on a T-scale.
Another possibility for analysis is the approach used in the previous section – the comparison of the cut scores ($\boldsymbol{CM}_i$ and $\boldsymbol{CM}_j$) obtained though the application of two different methods for obtaining the same cut score *X* (or *Y*), assuming that all other conditions (test, judgment and measurement scale) are equal. It is logical that such comparisons of the cut scores will be in pairs. The total number of these comparisons of combinations of methods is 252 and the summarized results are presented in Table 14. Table 14 shows the number of statistically significant differences ($|\boldsymbol{C}_x - \boldsymbol{C}_y|/\boldsymbol{SEM} > 2.77$) between the cut scores for each pair of methods above the diagonal, and below the diagonal are the mean values for this ratio. None of these mean values is over 2.77, which means that as a whole the different methods lead to similar cut scores.

Table 14 Significance of the differences between the different methods

Method		Number (%) of statistically significant differences						Total
		M1	**M2**	**M3**	**M4**	**M5**	**M6**	
Mean value $\lvert C_{Ni} - C_{Mj} \rvert / SEM$	**M1**	(n_{1j} = 48)	3 of 12 (25%)	2 of 12 (17%)	7 of 12 (58%)	1 of 6 (17%)	2 of 6 (33%)	**15 of 48 (31%)**
	M2	2.08	(n_{2j} = 48)	0 of 12 (0%)	1 of 12 (8%)	0 of 6 (0%)	1 of 6 (17%)	**5 of 48 (10%)**
	M3	1.97	0.22	(n_{3j} = 48)	1 of 12 (8%)	0 of 6 (0%)	1 of 6 (17%)	**4 of 48 (8%)**
	M4	2.69	1.09	1.20	(n_{4j} = 48)	0 of 6 (0%)	0 of 6 (0%)	**9 of 48 (19%)**
	M5	1.53	0.43	0.43	1.22	(n_{5j} = 48)	0 of 6 (0%)	**1 of 30 (3%)**
	M6	2.21	1.01	1.10	0.66	1.05	(n_{6j} = 48)	**4 of 30 (13%)**
Total		**2.16**	**1.03**	**1.04**	**1.48**	**0.93**	**1.21**	**1.21 (*n* = 252)**

As could be expected, the Basket procedure differed most from the others (with a mean value of difference of 2.16 from the other methods). Its cut scores are different from the ones obtained with the other methods in almost 1/3 of the cases (31%). Most often these statistical differences are in comparison with the ROC-curve method (M4, in 58% of the cases). The lowest and less frequently statistically significant are the differences between the Basket procedure (M1) on the one hand and the Cumulative Cluster method (M3) or the Item Mastery method (M5) on the other.

The results in Table 14 add additional support for the third hypothesis because they clearly show the closeness between the cut scores obtained by the Cumulative Compound method (M2) and the Cumulative Cluster method (M3) for which the mean ratio is the lowest ($|C_X - C_Y|/SEM$ = 0.22), and the number of the statistically significant differences between them is equal to zero.

Closest to these methods is the Item Mastery method (M5) for which the number of the statistically significant differences with them (M2 and M3) is also equal to zero. Actually this method (M5) shows the smallest number of statistically significant differences – only one, which is between this method and the Basket procedure (M1).

3.2.4 Classification consistency

3.2.4.1 Replicability of the classification decisions

Depending on the set cut scores, the examinees are classified in several classification levels (levels of competence) – one more than number of the cut scores – in correspondence with their test score (higher or lower than a given cut score).

One of the main criteria related to the quality of the set cut scores is whether these classification decisions will be identical in a potential replication of the testing with the same examinees. This criterion for the consistency of the classification decisions (*decision consistency*) is among the main aspects of the internal validity of the cut scores (Kane, 1994, pp. 445-446; Hambleton & Pitoniak, 2006, p. 460; Reckase, 2006, p. 6; Cizek & Bunch, 2007, pp. 60-61 and others) and can be considered analoguous to the term 'reliability' in criterion-referenced testing (Subkoviak, 1980; Berk, 1980; Hambleton & Slater; Haertel, 2006 and others).

The two most frequently used indices of classification consistency are the total relative share or percentage (***p***) of identical decisions in both classifications (Hambleton & Novick,

1973) and the coefficient of agreement – ***k*** (Swaminathan et al., 1974), which is a correction of ***p*** taking into account the probability for agreement by chance.
The numerical values of these indices depend mostly on:

- the characteristics of the test (number of items, reliability, error of measurement and frequency distribution of the test scores);
- the number of the cut scores and the classification groups into which the test score scale is divided through these cut scores;
- the difference between the mean value of the cut score and the different cut scores.

For instance, the index ***p*** usually takes higher values when the cut scores are located at the two ends of the frequency distribution, while the index ***k*** takes higher values when the cut scores are relatively close to the mean value of the test score, i.e. closer to the middle of the frequency distribution (Subkoviak, 1980, p. 153; Wan et al, 2007, p. 24).
In addition the values of the two indices ***p*** and ***k*** also depend on the number of classification groups – the more groups or categories, the more difficult it is to reach high values for ***p*** and ***k*** (Wan et al, 2007, p. 24).
The criterion for acceptable values of ***p*** (≥ 85%) and ***k*** (≥ 0.60), which Subkoviak (Subkoviak, 1988, pp. 52-53) suggests should be seen only as tentative, are appropriate only for two classification groups. In the current study, however, there are three classification levels of competence for all of the three tests because the two cut scores (***X*** and ***Y***) divide the test score scale into three intervals.
Concerning the psychometric characteristics of the concrete tests, their influence on classification stability is understandable because a high consistency of classification decisions cannot be expected for results obtained by a test with questionable quality.
The main limitation in determining the classification consistency for a given test and the respective cut scores is the necessity of testing the very same examinees twice with the same test, which is hardly very likely in reality. That is why numerous researchers (Huynh, 1976; Subkoviak, 1976; Breyer & Lewis, 1994; Livingston & Lewis, 1995; Brennan & Wan, 2004; Lee, 2005; and others) have worked to develop methods for the assessment of this consistency in a single administration of the test.
In the present study, the method of Livingston and Lewis (Livingston & Lewis, 1995) is used since the evaluations of classification consistency made on its basis are done with high precision (Wan et al., 2007, p. 25). Moreover, the method is appropriate for determining classification consistency with more than two classification categories and a polytomous scoring of the test items, which is not typical of the majority of the other methods.
For calculating the two indices of classification consistency using the method of Livingston and Lewis, the computer program *BB-Class.exe* (Brennan, R., 2004) was used. The values of these indices for the different methods for each test and measurement scale are presented in Table 15.

Table 15 Classification consistency

Index	Test	Raw test score				Z-scale					
		M1	M2	M3	M4	M1	M2	M3	M4	M5	M6
***p* (%)**	**T1**	87	87	87	89	86	88	88	90	88	87
	T2	82	81	81	82	81	85	86	86	85	87
	T3	78	79	77	87	83	81	81	83	80	82
k	**T1**	0.79	0.78	0.78	0.78	0.79	0.80	0.79	0.79	0.79	0.78
	T2	0.71	0.70	0.70	0.70	0.68	0.74	0.76	0.76	0.74	0.78
	T3	**0.53**	0.63	0.63	0.69	**0.48**	0.66	0.66	**0.48**	0.66	0.65

An analysis of the rows for ***p*** and ***k*** in Table 15 shows that the values of classification consistency for the different methods are quite close to each other, and for the Z-scale they are relatively higher compared to the values for the raw test score scale. This result is most probably due to the fact that in transforming the test score to the Z-scale the items take different weights, which leads to increasing the *effective test length*. The effective test length is a term introduced by Livingston and Lewis as *"the number of discrete, dichotomously scored, locally independent, equally difficult test items necessary to produce total scores having the same precision as the scores being used to classify the test taker"* (Livingston & Lewis, 1995, p. 180).

The comparison of the values by columns in Table 15 shows that the classification consistency for the third test, regardless of the method used for setting cut scores, is lower compared to the other tests. This result is most probably due to the specific characteristics of this test (T3) which has 10 to 12 items fewer than the other two tests and has the lowest reliability ($\alpha = 0.90$) among the three.

A comparison of the values of ***p*** and ***k*** for the different tests, methods and measurement scales with the criteria of Subkoviak (Subkoviak, 1988) shows that in the majority of cases (53%) the percentage of consistency (***p***) is higher than 85%, and its minimum value is 77% (raw test score – T3/M3). In interpreting these results, however, it should be taken into account that Subkoviak's criteria are relevant only for cases of dichotomous classifications. The expected values for classification consistency are higher for dichotomous classifications compared to classifications with more than two categories.

In confirmation of this, let us imagine just for a moment that for the raw test score of test T3/M3 instead of two, only one cut score exists and it is the cut score ***X***, set by the third method (M3). If the classification consistency is calculated, then the respective percentage of consistency will be 86%, i.e. 9% higher than the percentage of consistency with three categories (***p*** = 77%). In an analogous manner, if we use the second cut score (***Y***) instead of ***X*** in order to distribute the examinees into two classification categories (test score ≤ *Y* and test score >*Y*), then ***p*** will be equal to 87%.

The coefficient ***k*** is relatively less influenced by the number of classification categories, and it meets the "rule of thumb" criterion of Subkoviak (***k*** ≥ 0.60) in 27 out of a total of 30 cases (90%) and the three values below 0.60 are found with the third test.

The general conclusion is that the cut scores that are set using each one of the six methods lead to an acceptable level of classification consistency which depends mostly on the

psychometric characteristics of the particular test, but not on the method that was used for setting the cut scores themselves.
The second conclusion is that, in general, higher classification consistency is achieved for the cut scores that are set using the Z-scale, i.e. when IRT is used for calculating the test score and the items have different weights.

3.2.4.2 Equivalence of the classification decisions

In the previous section the following question was answered: what will be the degree of consistency of classification decisions made based on the same cut scores in repeated administrations of the same test with the same examinees?
However, another interesting question is: what will be the degree of consistency of the classification decisions in a single administration of the test, but using cut scores that were set using the different methods?
Since the different methods lead to different cut scores (see Section 3.3.3), it can be expected that these differences will also lead to some differences in the classification decisions. The question, however, is how large will these differences be in the classification decisions for two separate methods for setting cut scores. The answer to this question will also answer the question of how equivalent the different classification decisions are and to what extent the different methods are comparable to each other.
As a beginning, let us look at the frequency distribution of the classification decisions for Test 1 (n = 900), which were made using the test score on the Z-scale with the cut scores that were set using the Basket procedure (M1) and the ROC-curve method (M4). The reason for this choice is that, on the one hand, the cut scores in these two methods are considerably different in most cases (Table 14), and on the other hand the classification consistency for both methods in this particular case is very high and has exactly the same value (Table 15: k = 0.79).
Figure 36 shows the frequency distribution of the classification decisions that were made using the two methods (M1 and M2) for setting cut scores and the cross-tabulation in Table 16 shows the degree of consistency for both classifications.

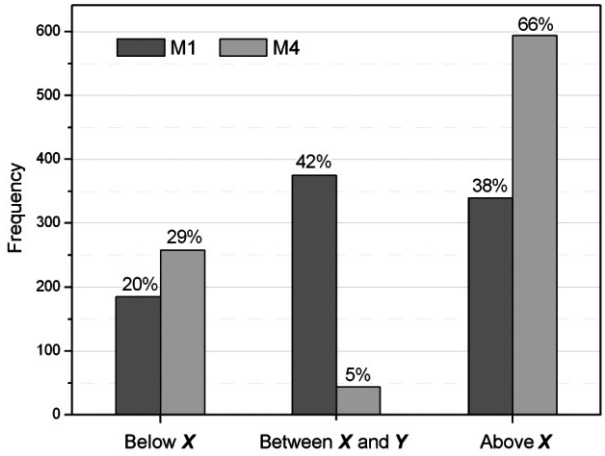


Figure 36 Frequency distribution of the classification decisions (T1Z)

Table 16 Classification consistency (T1Z: M1 and M4)

M1 \ M4	Below X	Between X and Y	Above Y	Total
Below X	**185**	0	0	185
Between X and Y	73	**44**	258	375
Above Y	0	0	**340**	340
Total	**258**	**44**	**598**	**900**
p = 63%; k = 0.45				

As expected, the two frequency distributions differ significantly not only in terms of the percentage of examinees classified at the three levels of competence, but also by their shape – the distribution for the Basket procedure (M1) is bell-shaped, while for the ROC-curve method (M4) it is U-shaped. This is due to the fact that, for the ROC-curve method, the two cut scores (X and Y) are considerably closer to each other, which leads to a lower number of examinees that are classified in the middle category (between X and Y). Moreover, only 44 of the total of 375 examinees (12%) classified by M1 as belonging to this level (between X and Y) are placed at the same level using M4. The rest of the examinees are classified at the lower or the higher levels of competence. In total, the percentage of identical classification decisions is 63% and both this percentage and the coefficient $k = 0.45$ are low enough to definitely reject the assumption of equivalence of the cut scores obtained using the two methods (M1 and M4) in this concrete case.
Concerning comparisons across all of the methods, Appendix 7 contains the indices for classification consistency (*p* and *k*) for each pair of methods for each test and measurement scale. For each of the tests, the corresponding values of ***p*** are provided **above** the main diagonal, and the values for ***k*** are provided **below** it. All values which go beyond Subkoviak's criteria (***p*** ≥ 85% and ***k*** ≥ 0.60) are in bold.
The comparison between the classification decisions made using the two different measurement scales (raw test score or Z-scale) is not quite correct because there is no

mutual equivalent correspondence. That is why the indices for these comparisons in Appendix 7 are presented with a grey font.[18]

From the table in Appendix 7, it can be seen that the percentage of correspondence between the classification decisions for the Compound Cumulative method (M2) and the Cumulative Cluster method (M3) varied between 91% and 100%. These values were lowest (91% and 95%) for the third test, for which also the lowest indices for classification consistency were found in terms of the stability of the decisions for the different methods (Table 15). Bearing in mind also that all values of *k* for this pair of methods are above 0.86, it can be concluded that the cut scores that were set using these two methods lead to equivalent classification decisions, which is another result **supporting the third hypothesis of the current study**.

To the group of these two equivalent methods (M2 and M3) can be added the Item Mastery method (M5), whose percentage of identical classification decisions with the Cumulative Compound method (M2) and the Cumulative Cluster method (M3) varied between 92% and 100%.

On the other hand, for the three comparisons of the classification decisions obtained using the ROC-curve method (M4) and the Level Characteristic Curves method (M6), the percentage of identical decisions was above 86%, and for Test 2 it reached 97% and *k* varied between 0.74 and 0.96 and also passed the rule of thumb minimum of 0.60. It can be argued that the application of the ROC-curve method (M4) and the Level Characteristic Curves method (M6) leads to equivalent classification decisions.

Concerning the Basket procedure (M1), from a total of 14 comparisons, there was not even one for which the percentage of identical decisions was higher than 85%. The highest value of the percentage of consistency was equal to 82% and related to the comparison between M1 and M5 for Test 1. This result is new, additional support for the **second research hypothesis**, since the cut scores obtained with this method led to classification decisions which are not equivalent to the classification decisions derived from the other five methods.

The results in Appendix 7 can be viewed as *proximity matrices* – two for each test depending on the measurement scale which is used for representing the test results. Each one of these matrices shows the degree of proximity between the pairs of the same variables – four variables in total (M1, M2, M3 and M4) for the raw test score and six variables (M1, M2, M3, M4, M5 and M6) for the Z-scale. One way of summarizing and visualizing the results from these proximity matrices is through application of *multidimensional scaling* – MDS.

Multidimensional scaling is a set of methods which transform one or several proximity matrices between pairs of stimuli into a map. On this map, each stimulus is represented as a point in the Euclidean space in a way that the distances between the points correspond

18 Editors' note: We wish to point out that the two methods refer to the same data and students are classified into one of three categories which have the same intended meaning. Therefore, it could be argued that differences are due only to differences in method of standard setting and the interpretation is the same as for other cases.

maximally[19] to the degree of proximity between the stimuli defined by the proximity matrices (Young & Harris, 1993; Groenen, & van de Velden, 2005; Takane, 2006). In its nature, multidimensional scaling is similar to factor analysis. However, MDS allows for the analysis of several proximity matrices at the same time. In addition, limitations concerning the structure and the characteristics of the input data are considerably fewer in MDS, which broadens its range of applications.

As a measure of proximity (d_{ij}) between a pair of stimuli (methods for setting cut scores), use was made of the relative share of inconsistent classification decisions ($d_{ij} = 1 - p_{ij}/100$), where p_{ij} is the corresponding percentage of identical decisions, presented in Appendix 7.

Since in the present study three different sources of data are analyzed (Test 1, Test 2 and Test 3), instead of the classical multidimensional scaling (CMDS), replicated multidimensional scaling (RMDS) was used, using SPSS and the PROXSCAL algorithm.

The analysis was conducted twice – separately for each of the two measurement scales (raw test score or Z-scale). In the first analysis (using raw test scores) only the first four methods (M1, M2, M3 and M4) were included since only they are applicable for setting cut scores using raw test scores. In the second analysis (Z-scale), however, all the six methods for setting cut scores were included.

The choice of concrete transformation of the measures of proximity for the different methods was made on the basis of the analysis of the degree of correspondence between the theoretical model and the empirical data for the different transformations. These analyses showed that the highest degree of correspondence for both scales (raw test score or Z-scale) was achieved by *ordinal* transformation through which the order of the methods is maintained, depending on the degree of proximity between them.

The choice of the number of dimensions was also dependent on the degree of correspondence between the chosen theoretical model and the empirical data. Table 17 shows the respective indices for this correspondence for each of the two analyses, and Figure 37 shows the change in the degree of stress[20] depending on the number of dimensions for the two analyses.

One of the main criteria for the quality of the results of multidimensional scaling is how good is the approximation of the real distances between the stimuli in the multidimensional space that was chosen. An index of this is the so called **stress-index**, for which different calculation formulae exist (Groenen, & van de Velden, 2005, p. 1291). Irrespective of the formula applied for the calculation of this index, the goal is its minimization. The closer to zero the value of the stress-index is, the closer to the reality is the representation of the stimuli in the multidimensional space that was chosen.

As Table 17 shows, the values for the different variations of the stress-index and in general the values of the normalized stress, which the PROXSCAL algorithm strives to minimize,

19 Editors' note: In MDS literature, the term proximity is used to indicate either a measure of similarity or a measure of dissimilarity. The author has clearly used the term as meaning dissimilarity..

20 Editors' note:The term 'stress' is a technical term from MDS literature. The author gives some explanation further down.

were very low (<0.04) regardless of the dimensionality in both analyses (for the raw test score and the Z-scale).

Table 17 Multidimensional scaling: Indices for the degree of correspondence between the theoretical model and the empirical data

Indices	**Number of dimensions**							
	Raw test score			**Z-scale**				
	1	**2**	**3**	**1**	**2**	**3**	**4**	**5**
Normalized stress	0,0312	**0,0005**	0,0004	0,0190	**0,0087**	0,0064	0,0045	0,0007
Stress I	0.1768	**0.0216**	0.0188	0.1378	**0.0931**	0.0798	0.0674	0.0264
Stress II	0.3720	**0.0574**	0.1348	0.2563	**0.2074**	0.2291	0.3234	0.2254
S – stress	0.1105	**0.0019**	0.0015	0.0494	**0.0186**	0.0150	0.0133	0.0027
% of the explained dispersion (DAF)	96.88	**99.95**	99.96	98.10	**99.13**	99.36	99.55	99.93

In the last row of Table 17 another index for the quality of the chosen multidimensional decision is presented. It shows what percentage of the dispersion in the original data can be explained through the multidimensional decision that was chosen. In contrast to the stress-index, the higher the values of this index, the better is the decision. The data in Table 17 shows that this index was rather high (> 96%) regardless of the number of the dimensions of the measurement scale.
Although it is close enough to the original proximity matrices, the one-dimensional decision is not the optimal one as Figure 37 shows. In both analyses (for the raw test score and for the Z-scale), the optimal decision is the two-dimensional Euclidean space. The reason for this is that the stress-index decreases significantly in the transition from a one-dimensional towards a two-dimensional decision in both analyses, while in the transition towards a three-dimensional decision the decrease of stress is lower, and for the raw test score it is almost equal to zero.

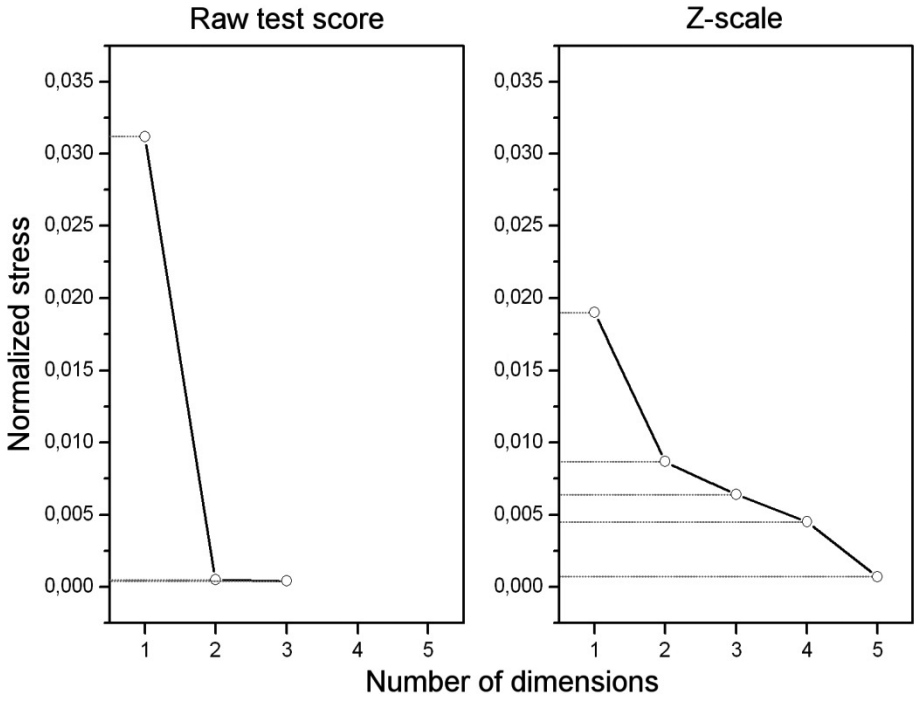


Figure 37 Multidimensional scaling: Changes in the stress-index depending on dimensionality

The graphical representation of the methods for setting cut scores in two-dimensional Euclidean space through ordinal transformation of the proximity between them is shown in Figure 38, separately for each one of the two measurement scales (raw test score and Z-scale).

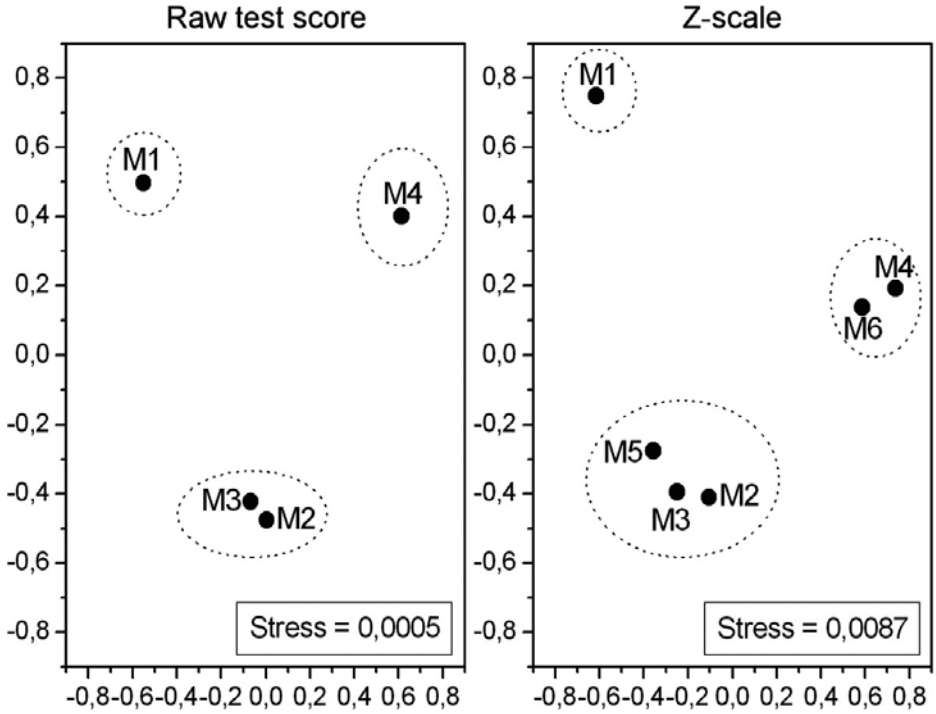


Figure 38 Positioning of the methods in the two-dimensional space in correspondence with the equivalence of the classification decisions

As both graphs in Figure 38 show, the location of the methods for setting cut scores in practice is the same regardless of whether they were applied to the test results expressed as raw test scores or on the Z-scale. The graphs in Figure 38 are in fact one additional strong confirmation of two of the three research hypotheses, namely:

- the Basket procedure (M1) differs from the other five methods (hypothesis II).
- the closest to each other in degree of consistency of the classification decisions are the Cumulative Compound method (M2) and Cumulative Cluster method (M3) – hypothesis III. The Item Mastery method (M5) can be added to this group.
- the other two methods – the ROC-curve method (M4) and the Level Characteristic Curve method (M6) form a distinct group, and the proximity between them is almost the same as between the Cumulative Cluster (M3) and the Item Mastery (M5) methods.

The interpretation of the two dimensions in multidimensional scaling is subjective, just as in factor analysis, and in most cases not straightforward or simple. In this particular case, however, the first (horizontal) dimension has a clear and logical interpretation – it can be interpreted in the context of the distance between two cut scores (*X* and *Y*). The lower the values of the first coordinate of the point corresponding to a given method, the bigger is the distance between the cut scores that were set using the method. This interpretation is consistent with the results of the classification consistency between the methods, and also with the results from the analysis of the significance of the differences between the two

cut scores for the different methods. For example, the Basket procedure (M1) is located furthest to the left and the distance between the two cut scores is bigger for this method, while the ROC-curve method for which the two cut scores are closest to each other is rightmost in the two-dimensional Euclidean space (Figure 38).
For the second (vertical) dimension, such a clear and logical interpretation, unfortunately, does not exist. One possible explanation of the lack of such an interpretation is the potential presence of the so-called *simplex-structure* of the six methods. The simplex-structure is typical of variables (stimuli) which are similar to each other in their nature, but are different in the degree of manifestation of some specific characteristic (difficulty, time, power, etc.). If a given group of stimuli for which a simplex-structure is visible is presented in two-dimensional space, the configuration of the points takes the shape of a horseshoe (Figure 39). In turn, the matrix of the distances between the separate points (Table 18) in the simplex-structure is characterized by the fact that the values for the separate cells are minimal around the main diagonal and the further away from the main diagonal a given cell is, the higher is the distance between the respective stimuli (Joreskog, 1978; de Leeuw & Michailidis, 1999). Using this matrix of the distances, the six methods can be arranged, depending on the magnitude of the difference between the two cut scores, in the following order: M1 > M5 > M3 > M2 > M6 > M4.

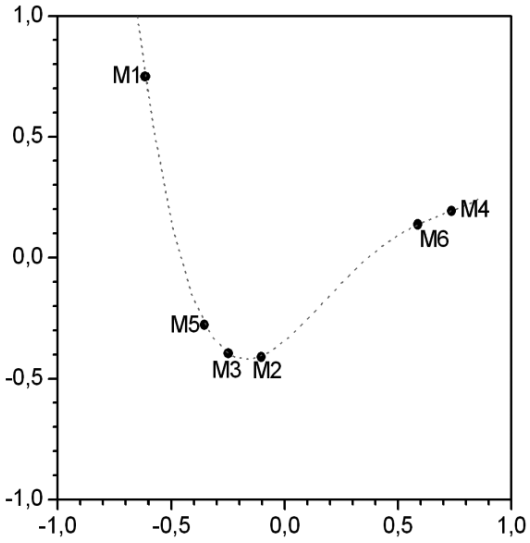


Figure 39 Multidimensional scaling (Z-scale)

Table 18 Matrix of the distances between the methods

	M1	M5	M3	M2	M6	M4
M1	**0.000**	1.057	1.200	1.266	1.349	1.461
M5	1.057	**0.000**	0.159	0.284	1.028	1.187
M3	1.200	0.159	**0.000**	0.144	0.989	1.146
M2	1.266	0.284	0.144	**0.000**	0.881	1.034
M6	1.349	1.028	0.989	0.881	**0.000**	0.160
M4	1.461	1.187	1.146	1.034	0.160	**0.000**

Since the simplex structure is one-dimensional in practice, this explains why the second dimension has no clear and meaningful interpretation.
In other words, the degree of equivalence between the classification decisions that were made using the different methods for setting cut scores depends mainly on the distance between two sequential cut scores (*X* and *Y*). If this distance is approximately equal for the compared pair of methods, then the classification decisions made on their basis are similar, i.e. equivalent as well.

3.3 Comparative analysis of the methods

3.3.1 Criteria for quality

Berk (Berk, 1986) was among the first who applied a system of criteria for a comparative analysis and evaluation of the quality of the different methods for setting cut scores.
The system suggested by Berk included 10 criteria consolidated in two main groups: technical adequacy and practicability. As Berk himself admitted, this system is incomplete since it misses criteria related to the quality of the judgments and linking of the judgments with the empirical data (Berk, 1986, pp. 144-145).
Van der Linden (van der Linden, 1994) suggested a system of six criteria (*explicitness, efficiency, unbiasedness, consistency, feasibility, robustness*). These six criteria are related mostly to the technical adequacy (*efficiency, unbiasedness, robustness*) and practicability (*explicitness, feasibility*) of the methods, but also take into account the quality of the judgments (*consistency*).
In turn Norcini and Shea (Norcini & Shea, 1997, p. 43) considered that in order to be accepted by a broad audience as adequate and realistic, the cut scores have to be set using a method that (a) produces absolute and not relative assessment standards; (b) is based on linking the results from informed expert judgment with the empirical data; (c) demonstrates due diligence, and (d) is supported by a body of research on its validity.
Norcini and Shea were not the first (Livingston & Zieky, 1982; Berk, 1986; Jaeger, 1990; Mehrens, 1994; Norcini, 1994 and others) who stressed the necessity of linking the subjective judgment with the available objective empirical data in setting the cut scores. They, however, were the first to raise this as an important quality criterion for the methods.
In his survey of 14 methods for setting cut scores, among which is also the Item Mastery method (M5), Reckase (Reckase, 2000-a, pp. 50-53) used a system of four main criteria as follows:

- Minimal level of distortion in converting judgments to a standard;

- Moderate to low cognitive complexity of the judgment tasks;
- Acceptable standard error of estimate for the cut scores;
- Replicable process for conducting the standard setting study with other groups of judges.

The first criterion defined by Reckase for minimal distortion of the cut score overlaps, to a large extent, with van der Linden's criterion for *unbiasedness*, although there are some differences in their interpretations of '*true* cut score'.

However, Reckase was the first who raised the cognitive complexity of the judgment task to the status of a criterion, even though the high cognitive complexity of the judgment task for the most widespread method – Angoff's probability method – was a major stimulus for developing new methods for setting cut scores. The judgmental task for the methods presented in this study has a cognitive complexity comparable to the lowest level of complexity identified by Reckase in a study of 14 methods (Reckase, 2000-a, p. 54).

Concerning the third and the fourth criteria, they overlap to a large extent with two of van der Linden's criteria (*efficiency* and *explicitness*), as his last criterion (guarantees for a potential replication or *explicitness*) is purely procedural and concerns not so much the method itself but its application for setting cut scores in a particular situation and good documention to make replication possible.

As this short overview shows, different scholars emphasize different criteria in evaluating the quality of the methods for setting cut scores, although there is some overlap among them. One of the reasons for the lack of complete consistency between the different systems of criteria is that the main factor on which the choice of a method for setting cut scores depends is the particular situation. The specific characteristics of the test, and the available resources, time and expertise are of prime importance in choosing a given method. That is why the choice of criteria will always depend on the particular situation, or as Mehrens points out, *"recognise that context matters in choosing standard-setting methods"* (Mehrens, 1994, p. 222).

Although some differences in choosing the criteria exist, there is common agreement (Kane, 1994, 1998, 2001; Hambleton & Pitoniak, 2006, Cizek & Bunch, 2007 and others) concerning their classification. The accepted classification scheme is based on the three main aspects of validity (procedural, internal and external), applicable to the cut scores and the methods used for establishing them.

If we apply this classification scheme to specific criteria presented earlier, we can see that the majority of methods fall into the first two categories – of procedural and internal validity with a slight predominance of internal validity over external validity. The reason for this predominance is that the first necessary condition for adequate and realistic cut scores is the presence of enough evidence to support the internal validity of the method being used. If such evidence is not available, no matter how strictly the procedure was applied for setting cut scores, the final results will be questionable.

The main reason for the slight neglect of external validity is that its testing requires the presence of an external criterion (in this case – another method for setting cut scores) whose validity also has to be confirmed. Moreover, the results from the present research, as well as many other studies, undoubtedly show that different methods applied to the same test situation can lead to different cut scores (Jaeger, 1989; Mehrens, 1994;

Bontempo et al, 1998 and others). Moreover, the differences are present even when there is enough evidence for the validity of the different methods.
Bearing in mind the existing criteria for the quality of the methods for setting cut scores, as well as the specificity of the particular situation, the following system of six criteria will be used for the comparative analysis of the six methods that are the subject of the present study, namely:

Criterion I *Range of application*

This criterion is **procedural** and takes into account only how broad the range of application of a given method is regarding the format of the test items, the scoring scheme and the approach used for the analysis and presentation of the test results (classical test theory or IRT).
Although this criterion is present in none of the aforementioned systems, it is of crucial importance in choosing the particular method for setting cut scores in a given situation because if the method is not applicable in this situation, its quality is irrelevant.

Criterion II *Statistical complexity*

This is another **procedural** criterion, which concerns the degree of complexity of the statistical procedures that are used for setting cut scores and whether additional software is required for their implementation.
This criterion is close to one of the four criteria for practicability suggested by Berk (Berk, 1986, pp. 143-144), according to which the *"...statistical methods used to arrive at the final standard should be easy to compute with either a calculator or an available statistical program on a microcomputer or mainframe"*.
Although the complexity of the method is not directly related to its quality in terms of validity, the simpler a given method is the broader the range of its application will be, which is the reason for adding this criterion.

Criterion III *Consistency with the empirical data*

This criterion is also **procedural** and to a large extent identical with the criterion of Norcini and Shea (Norcini & Shea, 1997) for setting the cut scores through aggregation of the empirical data with the results of the judgment.
Formulating this criterion, however, Norcini and Shea had in mind only providing empirical data to the experts themselves, which suggests a multi-stage judgment.
However, for greater economy in terms of time, for most of the methods included in the present research a different approach was used. This approach uses only a single-stage judgment without access to empirical data, but the final cut scores are set through an aggregation of the results from the judgment with the available empirical data.
Although procedural, this criterion is directly related to the internal validity of the methods for setting cut scores because it is related to the degree of consistency between the judgment and the empirical data.

Criterion IV *Misplacement of the cut score*

This criterion is the first related to the **internal validity** of the methods. It is completely identical with the first criterion suggested by Reckase (Reckase, 2000-a, pp. 50-51) for minimal distortion of the cut score in the transformation of the judgments into a performance standard, compared with the ideal, expected standards in which it is assumed the judge has all the empirical data, has understood completely the judgment task and

knows what the cut score is that he or she wants to set in advance.
Although it is listed in the fourth position, this criterion is one of the most important because if the method does not allow replication of a cut score that is known in advance through judgment that is completely consistent with the empirical data, then obviously there is a systematic error that is in the method itself.

Criterion V *Standard error of the cut scores*

This is the second most important criterion directly related to the **internal validity** of the method and the cut scores that were set using it. The requirement for a minimal standard error of the cut scores, which is a guarantee for replicability of the cut scores in a potential replication of the procedure in one way or another, is included in all systems of criteria for the quality of the methods.
The reason for its importance is that the lower the standard error the higher is the probability of a replication of the same cut scores through replication of the judgment with a different group of judges. This, in turn, will be a confirmation of the replicability of the cut scores.

Criterion VI *Significance of the differences between two consecutive cut scores*

This criterion is also directly related to the **internal validity** of the method and more specifically to the classification consistency of the decisions. If the difference between two consecutive cut scores (*X* and *Y*) is not statistically significant, then the probability that a given individual will be classified in a different way in two consecutive testing events becomes very high, especially if his or her result is close to the two cut scores.
Considering its importance, it is strange that this criterion has no counterpart in any of the existing systems of criteria for quality. Most probably this is due to the fact that setting more than one cut score, although frequent in reality, still has not attracted enough attention in the scientific community. Another possible explanation is that in cases of such proximity between the cut scores, in practice they may have been corrected through shifting (±*SEM*) of the cut scores without documenting or publishing this correction.
As can be seen from this description, this system of six criteria does not include any related to the external validity of the methods for setting cut scores. There are three reasons for this.
First, the analysis of external validity is impossible without the availability of an external, valid criterion. This is a serious problem in the field of setting cut scores because even the most widespread methods lead to different cut scores.
Second, since in this study each of the methods can play the role of an external criterion for the rest, the comparative analysis of the empirical data (Section 3.2) can also be regarded as an analysis of the external validity of the different methods.
Third, the methods which are the subject of the present research are relatively new and not supported by a sufficient number of scientific studies. That is why the present study aims to analyze their internal validity. Based on the present results, additional studies can be conducted in future to analyze the external validity of those methods that have shown to possess good quality in the present study.
In the system of six criteria presented here, there are also none that are related to the quality of the judgment itself (internal consistency and consistency with the empirical data), despite the fact that these are usually among the most commonly listed criteria.
The reason for this is not an underestimation of the importance of these criteria, but stems

from the fact that the cut scores for all methods in this study were set using the same type of judgments. In other words, although it is of extremely great importance, in this concrete case the quality of the judgment task itself does not influence the quality of the methods because they all were applied using the same judgments.
The system of six criteria was used in a comparative analysis of the quality of the different methods which are the subject of the present study. The results from this comparative analysis are presented in the next section.

3.3.2 Evaluation of the quality of the methods

In evaluating the quality of the six methods for setting cut scores in terms of each particular criterion, a pair comparison for all possible pairs (36) of methods was used. The matrices of these comparisons (one for each criterion) are presented in Appendix 8. The separate cells of the matrices ($\boldsymbol{m_{ij}}$) can take the following values:

- **0** – if method ***j*** excels method ***i*** for the given criterion ($M_i<M_j$);
- **1** – if method ***i*** excels method ***j*** for the given criterion ($M_i>M_j$);
- **0.5** – if both methods are equivalent for the given criterion ($M_i=M_j$).

This way of coding, since each method is equivalent to itself, regardless of the criterion of comparison, sets the values on the main diagonal for each matrix to be equal to 0.5.
In addition, if $\boldsymbol{m_{ij}} = 0$, then $\boldsymbol{m_{ji}} = 1$ and vice versa because if $M_i<M_j$, then $M_j>M_i$. For example, for the first criterion (range of application), the ROC-curve method (M4) exceeds the Item Mastery method (M5) because it is applicable to all kinds of scales, while the Item Mastery method (M5) can be applied only to tests developed using IRT. Since M4>M5, then $m_{54} = 1$ and $m_{54} = 0$ in the matrix corresponding to the first criterion.
For each single matrix, the sum by rows which is provided in the last column (Σ), defines the rank order of the given method concerning the respective criterion. The larger this sum (max = 5.5), the better is the given method concerning the respective criterion.
Concerning the first criterion (*range of application*), the six methods are divided into two groups. The first group comprises the methods which are applicable to tests regardless of whether they have been developed using classical test theory or IRT. This group comprises the first four methods (M1, M2, M3 and M4). The second group of methods comprises only those which are applicable only to tests developed using IRT – the Item Mastery method (M5) and the Level Characteristic Curves method (M6).
Concerning *statistical complexity* (Criterion II), the simplest is the Basket procedure (M1) followed by the Cumulative Compound method (M2). All other methods require the availability of additional software and need more "technical" time for setting the cut scores. The cut scores for all methods, except the first (M1) and the last (M6), are set by aggregating and comparing the information from the judgment data with the empirical data. In the Basket procedure (M1) and the Level Characteristic Curves method (M6), the cut scores are set only using judgments and this is the reason why these two methods have lower indices for the third criterion (*consistency with the empirical data*).
Some *misplacement of the cut scores* (Criterion IV) is found for all methods and this is due to the formulation of the judgment task itself. The method with the lowest distortion is the Level Characteristic Curves method (M6), which will show perfect agreement between expected and obtained cut scores if the judgment is in absolute agreement with the

empirical data. The highest misplacement is found with the Basket procedure (M1) and the ROC-curve method (M4), and this is the reason they occupy the last positions for this criterion.

Supporting the first hypothesis of the study, the empirical data (Section 3.2.2.2) show that the *standard error of the cut scores* (Criterion V) is the lowest for the Cumulative Compound method (M2) and the Basket procedure (M1).

As a consequence of the misplacement of the cut scores towards the two ends of the test score interval, *the difference between two consecutive cut scores* (*X* and *Y*) reaches the maximum for the Basket procedure (M1). The Item Mastery method (M5) also shows the same tendency, and the differences between the cut scores seen for this method are statistically significant for the three tests. In the last position for this criterion (Criterion VI) are the ROC-curve (M4) and the Level Characteristic Curves (M6) methods. In these methods, for the majority of cases the differences between *X* and *Y* are not statistically significant (see Section 3.2.3.2).

The main conclusion from this comparative analysis of the methods for the different criteria is that each has advantages and disadvantages. This conclusion, although logical, is actually quite trivial because (a) it was already known during the description of the methods in Chapter Two and (b) does not answer the question which of the six methods is the most effective.

The answer to this question, which is the main goal of the study, will be provided by the analysis of the individual profiles of the different methods (Figure 40).The six radii of each graph correspond to the six criteria for evaluating the quality of the methods. The thick black curves outline the individual profile of each of the methods and connect the points on these radii which correspond to the rank position of the method for the respective criteria.

The area of each profile is directly proportional to the effectiveness of the respective method. The larger the area, the higher are the rank positions of the given method and it can be defined as more effective.

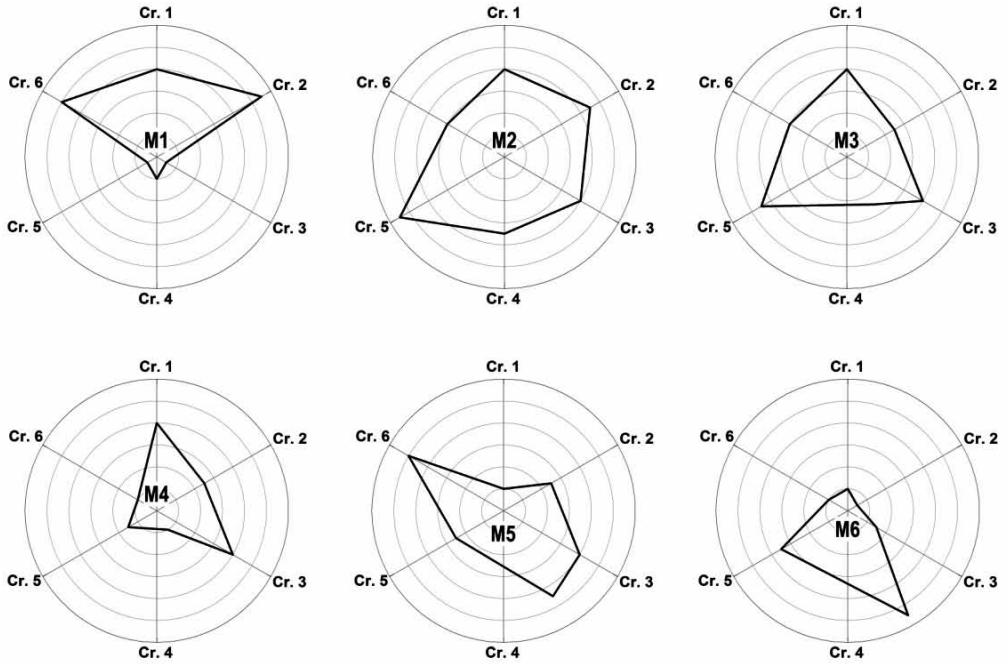


Figure 40 Individual profiles of the methods for setting cut scores for the different criteria

The analysis of the profiles shows that the **Compound Cumulative method** (M2) has the maximum area and is accordingly **the most effective** of the six methods according to the system of criteria developed for assessing quality. This conclusion is fully consistent with the conclusion that can be drawn using the average rank of the methods. Irrespective of the measure of central tendency used, the average rank of the Compound Cumulative method is the highest (mean = 4.08; median = 4.00; mode = 4.00). Moreover, in contrast to some of the other methods (for example M1, M5, M6), which for some of the criteria are in the first position and for others in the last, this method (M2) shows a stable pattern of behaviour and for all criteria is among the first three methods. Based on this analysis, we can conclude that the question that was asked in the beginning of this study, *"Which is the most effective method?"* has a clear and unequivocal answer: **the Cumulative Compound method**.

The next two methods which, although having lower indices compared to M2, have characteristics that are good enough to be recommended for broader application are the Cumulative Cluster method (M3) and the Item Mastery method (M5). The main reason for this conclusion is that their limitations are mainly procedural (rather more complex and with a narrower range of application) but the results from the present study do not cast any doubt on their internal validity. Moreover, if the Cumulative Compound method (M2) is used as an external criterion, then their external validity (in terms of consistency of the classification decisions and degree of proximity of the cut scores obtained by the different methods) is also very high because, as was already found, the cut scores obtained by these three methods are in practice equivalent.

In turn, the use of the Basket procedure (M1) is not recommended for application, especially in cases when the obtained cut scores will be used in making important decisions concerning individuals. The reasons are:

a this method uses only judgment and no empirical data.
b the standard error of the cut scores is large enough to cast doubt on the ability to replicate these cut scores.
c there is considerable distortion of the cut scores in terms of underestimation of the lower and overestimation of the higher cut score.

The only area where this method can be applied, due to its simplicity and economy, is formative assessment where the larger standard error can be compensated with a larger number of measurements.

The most ineffective method, according to the criteria used in this study, are the ROC-curve (M4) and the Level Characteristic Curves (M6) methods. Their application is not recommended because the results from the present study cast doubt on their internal validity – statistically non-significant difference between the two cut scores, large standard error and distortion for M4, lack of consistency with the empirical data for M6. Although these two methods are not recommended for setting cut scores, the approaches used by them can be successfully used for analysis of the degree of consistency between the judgment data and the empirical data. The results from such an analysis can be used in turn as reference points in planning the training of the judges.

Conclusions and recommendations

The valid interpretation of the results from a criterion-referenced achievement test is determined by two major factors: good psychometric characteristics of the test and adequately set cut scores. As the results from the present study clearly demonstrate, the cut scores themselves will be influenced to a large extent by the choice of the particular standard-setting method used for setting them. That is why a leading criterion for choosing a particular method for setting cut scores for a given test from the current great variety of available methods (over 60) should be the degree of validity of the method itself. The analysis of the validity of six concrete, relatively new and not yet adequately explored, test-centered methods for setting cut scores was one of the main goals of the current study. Based on evidence that support their procedural and internal validity, it was possible to define: (a) which of these methods have the required qualities to be recommended for broader application and (b) which of these methods should not be used due to serious limitations from both theoretical and practical perspectives which cast doubt on the adequacy of the cut scores obtained with them.
Besides achieving the main goal (determining the most effective method), based on the analysis of the results from the current research, several **concrete conclusions** can be drawn, as follows:

I Different methods lead to different cut scores

This conclusion is not new at all – it only confirms the results of many other prior studies which have reached the same conclusion. What is new in this case is that different cut scores are obtained even when different methods are applied to the results from the same judgments.
Previous studies have suggested the main reason for differences in the cut scores may be the fact that the judgment task for the different methods was different and this subsequently leads to differences in the cognitive processes which underlie the judgment tasks and this is translated into differences in the final results derived from the different types of judgment task. The personal characteristics of the judges and the quality of their preliminary instruction are often cited as additional reasons.
In the present study, however, for each test, the same judgments from the same judges who had the same training for the setting of cut scores for all the six methods were used. That is why, without underestimating the other possible reasons for differences between the cut scores obtained using different methods, in this particular case we can identify two essential reasons for difference – (a) differences in the way of generalizing the results from the judgment and (b) differences in the way of aggregating the results from the judgment data with the empirical data.
The present research addressed six different possibilities for generalizing the results from the judgment data and their aggregation with the empirical data. This, however, does not exhaust the full list of possibilities which in reality is unlimited. That is why, along with the aforementioned conclusion, concrete **recommendations are also** given, namely:
Recommendation 1 *Every single application of cut scores in the interpretation of the test*

results has to be accompanied by technical documentation which must provide a detailed description of the particular method that was used and the way it was applied in the particular situation.

Recommendation 2 *The application of any modification of the way the data from the judgments is generalized and its aggregation with the empirical data has to be accompanied by an in-depth analysis of the validity of the suggested modification.*

II The Cumulative Compound method is optimal for setting cut scores among the six alternatives

This conclusion is directly related to achieving the main goal of the current research, namely determining which one of the six analyzed methods is the most effective one in terms of procedural and internal validity. Of course, determining the most effective method depends to a large extent on the selection of criteria which will be used for comparing the methods. In this particular case, in the development of a set of criteria for a comparative analysis of quality, both the existing criteria for evaluating quality and the context of the particular situation were taken into account, and at the same time a balance between the different aspects of validity was sought.

The Cumulative Compound method was determined to be the optimal method for setting cut scores because it provided the best combination across several criteria: range of application, simplicity of the calculations, consistency with the empirical data, relatively low degree of distortion and statistically significant differences between the different cut scores. In addition, the Cumulative Compound method is the method with the lowest standard error, which confirms the first hypothesis of the study.

Since, according to the criteria used for analysis in this study, the Cumulative Compound method is the method for which an optimal balance of the procedural and internal aspects of validity was achieved, it is logical that it should be preferred for broad application over the other methods evaluated. The next recommendation is related to this.

Recommendation 3 *Among the six analyzed methods for setting cut-off cores, the most appropriate for mass use is the Cumulative Compound method.*

Irrespective of the fact that the results from the present study supported unequivocally the superiority of the Cumulative Compound method over the other five, the analysis of its validity should not be regarded as final and definitive. It is necessary to conduct additional research related to the analysis of other aspects of validity that are related to the replicability of the cut scores. Studies should be done in which the method is applied to observe the effect of manipulating other conditions not tested in this study – for example when used with different groups of judges, different sub-sets of test items, etc. It is also necessary to conduct comparative studies related to determining the external validity of the method, which should use as the external criterion some methods that have been affirmed through established practice over time like Angoff's probability method and the examinee-centered methods using contrasting or borderline groups.

Recommendation 4 *The process of further validation of the Cumulative Compound method should be continued and it should be focused on the other indices of internal replicability and consistency, as well as the external validity of the method.*

III The Basket procedure differs significantly from the other five methods

This conclusion in practice confirms the second hypothesis and is based both on the comparative analysis of the cut scores obtained using the different methods and the analysis of the classification decisions and main characteristics of the separate methods. The main disadvantages of this method are as follows:

- lack of consistency with the empirical data;
- considerable distortion of the cut scores toward the ends of the interval in which the test results vary;
- large standard error of the cut scores.

Although the Basket procedure is among the most widespread methods in the area of foreign language testing in Europe at the moment, bearing in mind the disadvantages that were found, the only possible recommendation is:

Recommendation 5 *Due to the serious disadvantages related to internal validity, the application of the Basket procedure should be limited only to tests used for formative assessment.*

IV The Cumulative Compound, Cumulative Cluster and the Item Mastery methods produce commensurable cut scores

The equivalence of the obtained cut scores obtained using these three methods confirm the third hypothesis of the present research, and this can be used as an argument supporting the external validity of each.

The similarities between the cut scores obtained by the Cumulative Compound and the Cumulative Cluster methods can be easily explained and predicted due to the similarity of the process of setting the cut scores with these methods. The Item Mastery method, however, differs significantly in the way of setting the cut scores. That is why the fact that the final cut scores are statistically non-significant, irrespective of the different methodologies for setting them, can be interpreted as additional confirmation of their adequacy and correspondingly as evidence for the validity of the methods used for their setting.

The higher statistical complexity of these two methods and the more limited range of application of the Item Mastery method are the basis of the next recommendation.

Recommendation 6 *Due to the necessity of using additional statistical software, the Cumulative Cluster and Item Mastery method are appropriate mainly as secondary methods for the external validation of cut scores that have already been set.*

V The two methods that have the lowest quality, according to the adopted system of criteria, are the ROC-curve and the Level Characteristic Curves methods

The main disadvantage of these two methods is the statistically non-significant difference between the two cut scores, which makes them inappropriate for cases when the setting of more than one cut score is required.

Additional limitations for the ROC-curve method are the large misplacement of the cut scores as well as their large standard error. For the Level Characteristic Curves method, the limitations are the lack of consistency with the empirical data, statistical complexity and its limited range of application.

Irrespective of their disadvantages, however, the two methods offer additional possibilities for an in-depth analysis of the consistency of the judgment data with the empirical data, which can be used successfully for the validation of the cut scores obtained by some other method. The next recommendation is related to this.

Recommendation 7 *The ROC-curve method and the Level Characteristic Curves method are appropriate only for an analysis of the consistency between the judgment data and the empirical data, but not for setting cut scores.*

VI The Misplacement Index is the optimal index for the degree of consistency between the empirical data and the type of judgments used in this study

This additional conclusion is not related directly to the main goal of the present study. However, it is of great importance in the analysis of the validity of the cut scores obtained with a method that uses this kind of judgment task.

The degree of consistency between the judgments and the empirical data is one of the major aspects of the internal validity of the obtained cut scores and the lack of an appropriate quantitative index can be viewed as a serious methodological limitation of such judgments and the methods for setting cut scores that are related to it. That is why the new index of consistency (MPI), created especially for this purpose, can be viewed as an important methodological contribution in the area of criterion-referenced testing.

The main advantages of this index are its clear and logical interpretation, ease of calculation and the possibility for determining its value both for each single judge and each single test item. This characteristic of the misplacement index broadens significantly the opportunities for its application. One limitation of this index is that although simple, its manual calculation is labour-intensive, which logically leads to the next recommendation.

Recommendation 8 *With a view to the broader application of the misplacement index (MPI), in the analysis of the degree of consistency between the judgment data and the empirical data, it is necessary to develop a software program for its calculation and to ensure free access to it.*[21]

VII The methods for setting cut scores discussed in the present study are not appropriate for the Bulgarian education system at the present time

The main reason for this pessimistic conclusion is that the methods discussed in the current study are based on the supposition that there is a strict, clear, valid and widely accepted system that is rich in content with criteria that describe in detail what a given individual is supposed to know and be able to do if he or she is at a certain level of competence. Such a system of criteria for the six-level marking system in Bulgaria still does not exist. The so called "qualitative" mark (fail, average, good, very good and excellent) is rather a set of labels that have no concrete meaning.

That is why, in the Bulgarian context, the more appropriate methods are those that are based on the term *minimally competent candidate*. These methods can be applied without

21 Editors' note: EALTA Members will be notified in due time when and how they can access the programme.

any limitations in our circumstances, but require a definition of the term *minimally competent candidate* to be arrived at in advance for each level of competence and for each particular achievement test. Defining this term is a compulsory element of the training of the judges participating in the judgment task.

Unfortunately, the main problem of achievement testing in Bulgaria is not choosing which method for setting cut scores will be the best to use, but first creating a realization of the necessity of applying such methods. Realization of this necessity would mean the end of the age of innocence, according Zieky's classification (Zieky, 1994, p. 4) and the beginning of the age of awakening.

Since any recommendation in this direction would sound rather like wishful thinking, the chanting of empty slogans or a line from a song calling for Bulgarian national revival, the present study will refrain from making concrete recommendations in this respect. It would also probably be wise to note that usually after the age of awakening comes the age of disillusionment. This, it should be noted, is a cautionary tale which is by no means new: *"For in much wisdom is much grief; and he who increases knowledge increases sorrow."* (Book of Ecclesiastes, 1:18)!

References

Бижков, Г. *Теория и методика на дидактическите тестове.* С., Просвета, 1992.
[Bizhkov, G. *Theory and Methods of the Achievement Tests*. Sofia, Prosveta, 1992.]

Бижков, Г. *Педагого-психологическа диагностика,* С., Университетско издателство, 2003.
[Bizhkov, G. *Pedagogical and Psychological Diagnostics*. Sofia, University Press "St. Kliment Ohridski", 2003.]

Вындев, Д. *Записки по приложна статистика 2*, С., 2003, [23.07.2007], <http://www.fmi.uni-sofia.bg/fmi/statist/index.html>
[Vandev, D. *Notes in Applied Statistics 2*. Sofia, 2003, [23.07.2007], <http://www.fmi.uni-sofia.bg/fmi/statist/index.html>.]

ЛПИ, България. *Сертифициране: Методология на разработка – Създаване на изпити,* 2006, [08.03.2007], <http://www.lpi-bulgaria.org/certification/exam_creation.php>
[LPI, Bulgaria. *Certification: Methodology of Development – Creation of exams,* 2006, [08.03.2007], <http://www.lpi-bulgaria.org/certification/exam_creation.php>.]

Министерство на образованието и науката, *Наредба № 1 за учебно-изпитните програми за държавните зрелостни изпити,* Изм. и доп., ДВ, 47, 2004, [15.03.2007], <http://www.minedu.government.bg/ opencms/opencms/left_menu/documents/process/>
[Ministry of education and science, *Ordinance № 1for the Education-Examination Programs for the State Maturity Exams*. SN, 47, 2004, [15.03.2007], <http://www.minedu.government.bg/ opencms/opencms/left_menu/documents/process/>.]

Министерство на образованието и науката, *Наредба № 3 от 15.04.2003 г. за системата за оценяване.* Изм. и доп., ДВ, 65, 2005, [15.03.2007], <http://www.minedu.government.bg/ opencms/ opencms/left_menu/documents/process/>
[Ministry of education and science, *Ordinance №3 from 15.04.2003 for the System of Assessment*. Changed and complemented by SN, 65, 2005, [15.03.2007], <http://www.minedu.government.bg/opencms/ opencms/left_menu/documents/process/>.]

Министерство на образованието и науката, *Наредба за изменение и допълнение на Наредба № 1 от 2003 г. за учебно-изпитните програми за държавните зрелостни изпити,* Обн., ДВ, 98, 2006, [13.02.2007], <http://www.minedu.government.bg/opencms/opencms/ left_menu/ documents/process/>
[Ministry of education and science, *Ordinance for Modification and Complementation of the Ordinance №1 for the Education-Examination Programs for the State Maturity Exams from 2003*. SN, 98, 2006, [13.02.2007], <http://www.minedu.government.bg/opencms/ opencms/ left_menu/documents/process/>.]

Министерство на образованието и науката, *Българско гражданство*, [20.03.2007], <http://mon.bg/opencms/opencms/left_menu/citizenship/>
[Ministry of education and science, *Bulgarian Citisenship*, [20.03.2007], <http://mon.bg/opencms/opencms/left_menu/citizenship/>.]

Национална програма за развитие на училищното образование и предучилищното възпитание и подготовка (2006-2015 г.). // ДВ, 50, 2006, 2-17, [16.09.2006], <http://www.minedu.government.bg/ opencms/opencms/>
[National Program for Development of the School Education and Pre-school Education and Training (2006-2015). SN, 50, 2006, 2-17, [16.09.2006], <http://www.minedu.government.bg/ opencms/opencms/>.]

Националната агенция за професионално образование и обучение, *Методически указания за разработване на Държавни образователни изисквания за придобиване на квалификация по професии*, 2004, [05.05.2007], <http://www.navet.government.bg/Sections/Professional.htm>
[National Agency for Vocational Education and Training, *Methodical Directions for Developing State Educautional Requirements for Acquiring Professional Qualifications*, 2004, [05.05.2007], <http://www.navet.government.bg/Sections/Professional.htm>.]

Националната агенция за професионално образование и обучение, *Професионална квалификация: Държавни образователни изисквания, публикувани в Държавен вестник*, [05.05.2007], <http://www.navet.government.bg/navet_vq/doi_dv/Prof-doi.htm>
[National Agency for Vocational Education and Training, *Proffessional Qualification: State Educautional Requirements Published in State Newspaper*, [05.05.2007], <http://www.navet.government.bg/navet_vq/doi_dv/Prof-doi.htm>.]

Правила за използването на тестове на Международния тестов комитет (МТК), Българска версия, 2000, [16.03.2007], <http://www.psychology.mvr.bg/Professional_standarts/tests.htm>
[Rules for test use from the International Test Commission (MTK), Bulgarian Version, 2000, [16.03.2007], <http://www.psychology.mvr.bg/Professional_standarts/tests.htm>.]

Симидчиев, Г. Практически методи за оценяване на отговорите на учениците, подложени на тестове с множествен избор на отговор. Преход от тестов бал към шестобална система. // *Образование и професия*, 1996, 8 (12), 18-20.
[Simidchiev, G. Practical methods for scoring the student's responses to multiple choice and constructed response test items. Transition from raw test score into six-point marking scale. // *Education and occupation*, 1996, 8 (12), 18-20.]

Стоянова, Ф. Приложение на теорията на размитите множества в областта на критериално-ориентираното тестиране. // *Психология*, 1992, 1, 63-68.
[Stoyanova, F. Application of the fuzzy sets theory in the criterion-referenced testing. // *Psychology*, 1992, 1, 63-68.]

Стоянова, Ф. *Тестология за учители*. С., Атика, 1996-а.
[Stoyanova, F. *Test theory for teachers*. Sofia, Atika, 1996-a.]

Стоянова, Ф. Тест за училищна готовност: описание и анализ. // Диагностика на готовността на децата за училище, Колектив с р-тел Г. Бижков, София, УИ "Св. Кл. Охридски", 1996-б, 144-186.
[Stoyanova, F. School-Readiness Test: description and analysis. // Diagnostics of the school readiness of the children. G. Bizhkov (Ed.). Sofia, University Press "St. Kliment Ohridski", 1996-b, 144-186.

Стоянова, Ф. Приложение на еднопараметричния логистичен модел при конструирането и анализа на тестове. // Бижков, Г. *Теория и методика на дидактическите тестове*. София, Просвета, 1996-в, 284-309.
[Stoyanova, F. Application of the one-parameter logistic model in construction and analysis of the achievement tests. // Bizhkov, G. *Theory and Methods of the Achievement Tests*. Sofia, Prosveta, 1996-c, 284-309.]

Стоянова, Ф. Тест за разбиране при четене. // *Диагностика на грамотността (III част): Разбиране при четене*. Колектив с р-тел Г. Бижков, София, Университетско издателство „Св. Климент Охридски, 2004, 228-288.
[Stoyanova, F. Test for reading comprehension. // Bizhkov, G. (Ed.). *Diagnostics of the literacy (Vol. 3): Reading comprehension*. Sofia, University Press „St. Kliment Ohridski", 2004, 228-288.]

Съвет на Европа. *Обща европейска езикова рамка: Учене, преподаване и оценяване*. // Дж. Трим и др., Варна, РЕЛАКСА, 2006.
[Council of Europe. *CEFR for Languages: Learning, teaching and assessment*. // Trim, J. and others. Varna, RELAXA, 2006.]

Щетински, Д. Стандарти за създаване и оценка на нови тестове, адаптации или модификации на чуждестранни тестове в български условия и професионалното им прилагане. // *Българско списание по психология*, 2006, 2, 3-43.
[Shtetinski, D. Standards for creating and evaluating of new tests, adaptations and modifications of foreign tests in Bulgarian settings and their professional application. // *Bulgarian journal of psychology*, 2006, 2, 3-43.]

Ang, R. Use of the Jackknife Statistic to Evaluate Result Replicability. // *The Journal of General Psychology*, 1998. 125(3), 218-228.

Angoff, W. Scales, Norms and Equivalent Scores. // *Educational Measurement*. Ed. by R. L. Thorndike, Washington, 1971, 508-600.

Bay, L., & M. Nering. *A Demonstration of Using Person-Fit Statistics in Standard Setting*. Paper presented at the Annual Meeting of the American Educational Research Association. San Diego, CA, April 13-17, 1998.

Berk, R. A Consumers' Guide to Criterion-Referenced Test Reliability. // *Journal of Educational Measurement*, 1980, 17(4), 323-349.

Berk, R. A Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests. // *Review of Educational Research*, 1986, 56 (1), 137-172.

Berk, R. Something Old, Something New, Something Borrowed, a Lot to Do! // *Applied Measurement in Education*, 1995, 8 (1), 99-109.

Berk, R. Standard Setting: The Next Generation (Where Few Psychometricians Have Gone Before!). // *Applied Measurement in Education*, 1996, 9 (3), 215-235.

Beuk, C. A Method for Reaching a Compromise Between Absolute and Relative Standards in Examinations. // *Journal of Educational Measurement*, 1984, 21 (2), 147-152.

Biddle, R. How to Set Cutoff Scores for Knowledge Tests Used In Promotion, Training, Certification, and Licensing. // *Public Personnel Management*, 1993, 22 (1), 63-70.

Bontempo, Br. et al. *A Meta-Analytic Assessment of Empirical Differences in Standard Setting Procedures*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, 1998.

Bradshaw, J. & C. Kirkup. *Inventory of Language Certification in Europe. A Report to the European Commission Directorate General for Education and Culture*, NFER, 2006, [1.10.2006], <http://ec.europa.eu/education/policies/lang/doc/inventory.pdf>

Brandon, P. Conclusions About Frequently Studied Modified Angoff Standard-Setting Topics. // *Applied Measurement in Education*, 2004, 17 (1), 59-88

Brennan, R. & L. Wan. *Bootstrap Procedures for Estimating Decision Consistency for Single-administration Complex Assessments*. CASMA RR 7. Iowa, Center for Advanced Studies in Measurement and Assessment, 2004.

Brennan, R. *Manual for BB-CLASS: A Computer Program that uses the Beta-Binomial Model for Classification Consistency and Accuracy (Version 1.1)*, CASMA RR 9, Iowa, Center for Advanced Studies in Measurement and Assessment, 2004.

Breyer, F. & C. Lewis. P*ass-fail Reliability for Tests with Cut Scores: A Simplified Method*. ETS RR 94-39. Princeton, Educational Testing Service, 1994.

Cangelosi, J. Another Answer to the Cut score Question. // *Educational Measurement: Issues and Practices*, 1984, 3, 23-25.

Chang, L. Judgmental Item Analysis of the Nedelsky and Angoff Standard-Setting methods. // *Applied Measurement in Education*, 1999, 12 (2): 151–165.

Chang, L. et al. Does a Standard Reflect Minimal Competency of Examinees or Judge Competency? // *Applied Measurement in Education*, 1996, 9 (2), 161-173.

Chang, L. et al. Setting Standard and Detecting Inrajudge Inconsistency Using Interdependent Evaluation of Response Alternatives. // *Educational and Psychological Measurement*, 2004, 64 (5), 781-801.

Cizek, Gr. Reconsidering Standards and Criteria. // *Journal of Educational measurement*, 1993, 30 (2), 93-106.

Cizek, Gr. Standard-setting Guidelines. // *Educational Measurement: Issues and Practice*, 1996, 15 (1), 13-21.

Cizek, Gr. Conjectures on the Rise and Call of Standard Setting: An Introduction to Context and Practice. // S*tandard-setting: Concepts, methods, and perspectives*. Ed. by G. Cizek, Hillsdale, Erlbaum, 2001, 3-18.

Cizek, Gr. An NCME Instructional Module on Setting Performance Standards: Contemporary Methods. // *Educational Measurement: Issues and Practice*, 2004, 23 (4), 31-50.

Cizek, Gr. & M. Bunch. *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. London, SAGE Publications, 2007.

Clauser, Br. et al. Multivariate Generalizability Analysis of the Impact of Training and

Examinee Performance. // *Journal of Educational Measurement*, 2002, 39 (4), 269-290.

Cliff, N. Answering Ordinal Questions with Ordinal Data Using Ordinal Statistics. // *Multivariate Behavioral Research*, 1996, 31 (3), 331-350.

Cohen, A. et al. A Generalized Examinee-Centered Method for Setting Standards on Achievement Tests. // *Applied Measurement in Education*, 1999, 12 (4), 343–366.

Council of Europe. *CEFR for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press, 2001.

Council of Europe. *Relating Language Examinations to the CEFR for Languages: Learning, Teaching, Assessment (CEF): Manual – Preliminary Pilot Version*. Strasbourg: Language Policy Division, 2003.

Council of Europe. *Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the CEFR for Languages: Learning, Teaching, Assessment*. Strasbourg: Language Policy Division, 2004.

Council of Europe. *Survey on the use of the CEFR for Languages (CEFR): Synthesis of results*. 2005, [2.08.2006], <http://www.coe.int/t/dg4/linguistic/Source/Surveyresults.pdf>

CRESST – National Center for Research on Evaluation, *CRESST Assessment Glossary* // Standards, and Student Testing Web site, 1999, [12.12.2003], <http://www.cse.ucla.edu/CRESST/pages/glossary.htm>

De Champlain, A. et al. *Setting Test-Level Standards for a Performance Assessment of Physicians1 Clinical Skills: A Process Investigation*. Paper presented at the Anrlual Meeting of the AmericanEducational Research Association, San Diego, 1998.

De Gruijter, D. Compromise Models for Establishing Examination Standards. // *Journal of Educational Measurement*, 1985, 22 (4), 263-269.

De Leeuw, J. & G. Michailidis. *Graph Layout Techniques and Multidimensional Data Analysis*. Paper 1999010104, UCLA, Department of Statistics, 1999.

DeMauro, G. & D. Powers. *Logical Consistency of the Angoff Method of Standard setting*. RR-93-26, Princeton, Educational Testing Service, 1993.

Downing, St. et al. Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education. // *Teaching and Learning in Medicine*, 2006, 18 (1), 50–57.

Dylan, W. Meaning and Consequences in Standard Setting. // *Assessment in Education: Principles, Policy & Practice*, 1996, 3, (3), 287-308.

Ebel, R. L. *Essentials of Educational Measurement*. Englewood Cliffs, N.J., Prentice-Hall, 1972, 492-494.

Emrick, J. An Evaluation Model for Mastery Testing // *Journal of Educational Measurement*, 1971, 8 (4), 321-326.

Ercikan, K. & M. Julian. Classification Accuracy of Assigning Student Performance to Proficiency Levels: Guidelines for Assessment Design. // *Applied measurement in Education*, 2002, 15 (3), 269-294.

Fawcett, T. An Introduction to ROC Analysis. // *Pattern Recognition Letters*, 2006, 27(8), 861-874.

Fehrmann, M. et al. The Angoff Cutoff Score Method: The Impact of Frame-of-Reference Rater Training. // *Educational and Psychological Measurement*, 1991, 51, 857-872.

Feldt, L. et al. A Comparison of Five Methods for Estimating the Standard Error of Measurement at Specific Score Levels. // *Applied Psychological Measurement*, 1985, 9 (4), 351-361.

Ferdous, A. & B. Plake. Understanding the Factors that Influence Decisions of Panelists in a Standard-Setting Study. // *Applied Measurement in Education*, 2005, 18(3), 257-267.

Ferdous, A. et al. *Factors That Influence Judges' Decisions in an Angoff Standard Setting Study*. Paper presented at the 2006 annual meeting of the American Educational Research Association, San Francisco, CA.

Fisher, W. Reliability Statistics. // *Rasch Measurement Transaction*, 1992, 6:3, 238, [8.12.1999], <http://209.41.24.153/rmt/rmt63.htm>

Fredricks, Gr. & R. Nelsen. On the Relationship between Spearman's rho and Kendall's tau for Pairs of Continuous Random Variables. // *Journal of Statistical Planning and Inference*, 2007, 137, 2143 – 2150.

Gillaspy, J. *Evaluating Result Replicability: Better Alternatives to Significance Tests*. Paper presented at the Annual Meeting of the Southwest Educational Research Association, New Orleans, 1996.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. // *American Psychologist*, 1963, 18, 519-521.

Glass, G. Standards and criteria. // *Journal of Educational Measurement*, 1978, 15, (4), 237–261.

Goldman, A. *Knowledge in a Social World*, Oxfordq Clarendon Press, 1999.

Goodwin, L. Relations between Observed Item Difficulty Levels and Angoff Minimum Passing Levels for a Group of Borderline Examinees // *Applied Measurement in Education*, 1999, 12 (1), 13-28.

Green, B. *Setting Performance Standards*. Paper presented at MAPAC meeting, 2000.

Groenen, P. & M. van de Velden. Multidimensional Scaling. *Encyclopedia of Statistics in Behavioral Science*, Vol. II, Ed. by Br. Everitt & D. Howell, Chichester, Wiley, 2005, 1280-1289.

Grosse, M. & B. Wright. Setting, Evaluating, and Maintaining Certification Standards with the Rasch Model. // *Evaluation & The Health Profession*, 1986, 9 (3), 267-285.

Haertel, E. Reliability. *Educational Measurement*, NCME, R. Brennan (Ed.), Greenwood World Publishing, 2006, 65-110.

Haertel, E. & W. Lorie. Validating Standards-Based Test Score Interpretations. // *Measurement*, 2004, 2 (2), 61-103.

Haladyna, Th. & R. Hess. An Evaluation of Conjunctive and Compensatory Standard-Setting Strategies for test Decision. // *Educational Assessment*, 1999, 6 (2), 129-153.

Hambleton, R. On the Use of Cut scores with Criterion-Referenced Tests in Instructional Settings. // *Journal of Educational Measurement*, 1978, 15 (4), 277–289.

Hambleton, R. Test score validity and standard-setting methods. // *Criterion-referenced measurement: The state of the art*, Ed. By R. Berk, Baltimore, Johns Hopkins University Press, 1980, 80-123.

Hambleton, R. The Rise and Fall of Criterion-Referenced Measurement? // *Educational Measurement: Issues and Practice*, 1994, 13(4), 21-26.

Hambleton, R. Setting Performance Standards on Achievement Tests: Meeting the Requirements of Title I. // *Handbook for the Development of Performance Standards*, Ed. by L. N. Hansche, Washington, DC: Council of Chief State School Officers, 1998, 87-114.

Hambleton, R. et al. F*undamentals of Item Response Theory*. Newbury Park, Sage Press, 1991.

Hambleton, R. et al. Setting Performance Standards on Complex Educational Assessments. // *Applied Psychological Measurement*, 2000, 24 (4), 355–366.

Hambleton, R. & B. Plake. Using an Extended Angoff Procedure to Set Standards on Complex Performance Assessments. // *Applied Measurement in education*, 1995, 8 (1), 41-45.

Hambleton, R. & M. Novick. Toward an Integration of Theory and Method for Criterion-referenced Tests. // *Journal of Educational Measurement*, 1973, 10 (3), 159-170.

Hambleton, R. & M. Pitoniak. Setting Performance Standards. // *Educational Measurement*. NCME, R. Brennan (Ed.), Greenwood World Publishing, 2006, 433-470.
Hambleton, R. & Sh. Slater. Reliability of Credentialing Examinations and the Impact of Scoring Models and Standard-Setting Policies. // *Applied Measurement in Education*, 1997, 10 (1), 19-38.

Hansche, L. *Handbook for the Development of Performance Standards: Meeting the Requirements of Title I*. Washington, US Department of Education and the Council of Chief State School Officers, 1998.

Hanson, Br. & L. Bay. *Classifying Student Performance as a Method for Setting Achievement Levels for NAEP Writing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, 1999.

Harnish, D. & R. Linn. Analysis of Item Response Patterns: Questionable Test Data and Dissimilar Curriculum Practices. // *Journal of Educational Measurement*, 1981, 18 (3), 133-146.

Harvill, L. An NCME Instructional Module on Standard Error of Measurement. // *Educational Measurement: Issues and Practice*, 1991, 10 (2), 33–41.

Hattie, J. & G. Brown. *Standard Setting for asTTle Reading: A Comparison of Methods*, asTTle Technical Report 21, University of Auckland, 2003.

Hintze, J. & B. Silberglitt. A Longitudinal Examination of the Diagnostic Accuracy and Predictive Validity of R-CBM and High-Stakes Testing // *School Psychology Review*, 2005, 34(3), 372-386.

Hurtz, Gr. & N. Hertz. How Many Raters Should be Used for Establishing Cutoff Scores with the Angoff Method? A Generalizability Theory Study. // *Educational and Psychological Measurement*, 1999, 59 (6), 885-897.

Huynh, H. On the Reliability of Decisions in Domain-referenced Testing. // *Journal of Educational Measurement*, 1976, 13 (4), 253-264.

Impara, J. & B. Plake. Standard Setting: An Alternative Approach. // *Journal of Educational Measurement*, 1997, 34(4), 353-366.

Impara, J. & B. Plake. Teachers' Ability to Estimate Item Difficulty: A Test of the Assumptions in the Angoff Standard Setting Method. // *Journal of Educational Measurement*, 1998, 35 (1), 69-81.

Impara, J. & B. Plake. *A Comparison of Cut Scores Using Multiple Standard Setting Methods*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, 2000.

Jaeger, R. An Iterative Establishing Structured Judgment Process for Standards on Competency Tests: Theory and Application. // *Educational Evaluation and Policy Analysis*, 1982, 4 (4), 461-475.

Jaeger, R. Use and Effect of Caution Indices in Detecting Aberrant Patterns of Standard-Setting Judgments. // *Applied Measurement in Education*, 1988, 1 (1), 17-31.

Jaeger, R. Certification of student competence. // *Educational Measurement*, Ed. by R. Linn, Washington, DC: American Council on Education, 1989, 485-511.

Jaeger, R. Establishing standards for Teacher Certification Tests. // *Educational Measurement: Issues and Practices*, 1990, 9 (4), 15-20.

Jaeger, R. Selection of Judges for Standard-Setting. // *Educational Measurement: Issues and Practice*, 1991, 10 (2), 3-6.

Jaeger, R. Setting Performance Standards through Two-Stage Judgmental Policy Capturing. // *Applied Measurement in Education*, 1995, 8 (1), 15–40.

Jaeger, R. & C. Mills. An Integrated Judgment Procedure for Setting Standards on Complex Large-Scale Assessments. // *Standard-Setting: Concepts, Methods, and Perspectives*. Ed. by G. Cizek, Hillsdale NJ: Erlbaum, 2001, 313-338.

Janssen, R. et al. A Hierarchical IRT Model for Criterion-Referenced Measurement. // *Journal of Educational and Behavioral Statistics*, 2000, 25 (3), 285-306.

Jones, J. et al. *A Preliminary Investigation of the Direct Standard Setting Method*. Paper preseented at the Annual Meeting of the American Educational Research Association, New Orleans, 1988.

Joreskog, K. Structural analysis of covariance and correlation matrices. //*Psychometrika*, 1978, 43 (4), 443-477.

Kaftandjieva, F. Section B: Standard Setting. // Council of Europe. *Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the CEFR for Languages: Learning, Teaching, Assessment*. Strasbourg: Language Policy Division, 2004.

Kaftandjieva, F. et al, *DIALANG: A Manual for Standard Setting Procedure*, Unpublished, 2000.

Kaftandjieva, F. & N. Verhelst. *A New Standard Setting Method for Multiple Cut scores*. Paper presented at LTRC, Vancouver, 2000.

Kaftandjieva, F. & S. Takala. *Relating the Finnish Matriculation Examination English Test Results to the CEF Scales*. Paper presented at Helsinki Seminar on Linking Language Examinations to CEFR for Languages: Learning, Teaching, Assessment, June 31– July 2, 2002.

Kane, M. Validating the Performance Standards Associated With Passing Scores. // *Review of Educational Research*, 1994, 64 (3), 425-461.

Kane, M. Choosing Between Examinee-Centered and Test-Centered Standard-Setting Methods // *Educational Measurement*, 1998, 5 (3), 129-145.

Kane, M. So much remains the same: Conception and status of validation in setting standards. // *Setting performance standards: Concepts, methods, and perspectives*. Ed. by G. Cizek, Mahwah, Lawrence Erlbaum, 2001, 53-88.

Karabatsos, G. Comparing the Aberrant Response Detection Performance of Thirty-Six Person-Fit Statistics. // *Applied Measurement in Education*, 2003, 16 (4), 277-298.

Karantonis, A. & St. Sireci. The Bookmark Standard-Setting Method: A Literature Review. // *Educational Measurement: Issues and Practice*, 2006, 25 (1), 4-12.

Kier, Fr. *Ways to Explore the Replicability of Multivariate Results (Since Statistical Significance Testing Does Not)*. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, 1997.

Kingston, N. et al. Setting Performance Standards Using the Body of Work Method. // *Setting Performance Standards: Concepts, Methods, and Perspectives*. Ed. by G. J. Cizek, Mahwah, Lawrence Erlbaum, 2001, 219-248.

Lapata, M. Automatic Evaluation of Information Ordering: Kendall's Tau. // *Computational Linguistics*, 2006, 32 (4), 471-481.

Lee, W. et al. Estimating Consistency and Accuracy Indices for Multiple Classifications. // *Applied Psychological Measurement*, 2002, 26 (4), 412–432.

Lee, W. *Classification consistency under the compound multinomial model*. CASMA RR 13. Iowa, Center for Advanced Studies in Measurement and Assessment, 2005.

Leighton, J. et al. *The Attribute Hierarchy Model for Cognitive Assessment*. Paper Presented at the Annual Meeting of the National Council on Measurement in Education (NCME), New Orleans, Louisiana, 2002.

Linn, R. Criterion-Referenced Measurement: A Valuable Perspective Clouded by Surplus Meaning. // *Educational Measurement: Issues and Practice*, 1994, 14 (4), 12–14.
Linn, R. Linking Results of Distinct Assessments. // *Applied Measurement in Education*, 1993, 6 (1), 83-102.

Linn, R. Validating Inferences From National Assessment of Educational Progress Achievement-Level Reporting. // *Applied Measurement in Education*, 1998, 11 (1), 23-47.

Linn, R. *The Design and Evaluation of Educational Assessment and Accountability Systems*. CSE Technical Report 539. CREST/ University of Colorado, 2001.

Linn, R. Performance standards: Utility for different uses of assessments. // *Education Policy Analysis Archives*, 2003, 11 (31).

Livingston, S. *Translating Verbally Defined Proficiency Levels into Test Score Intervals*. Paper presented at the Annual meeting of NCME, Chicago, 1991.

Livingston, S. & Ch. Lewis. Estimating the Consistency and Accuracy of Classifications Based on Test Scores. // *Journal of Educational Measurement*, 1995, 32 (2), 179-197.

Livingston, S. & M. Zieky. *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton, NJ: ETS, 1982.

Loomis, S. & M. Bourque. From tradition to innovation: Standard-setting on the National Assessment of Educational Progress. // *Standard setting: Concepts,methods, and perspectives*. Ed. by G. Cizek, Mahwah NJ, Erlbaum, 2001, 175-218.

Lord, F. *Applications of Item Response Theory to Practical Testing Problems*, NJ, Lawrence Erlbaum Associates, 1980.

Lunz, M. *Performance Examinations: Technology for Analysis and Standard Setting*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Chicago, 1997.

Macready, G. & M. Dayton. The Use of Probabilistic Models in the Assessment of Mastery, // *Journal of Educational Statistics*, 1977, 2 (2), 99-120.

McGinty, D. & J. Neel. *Juidgmental Standard Setting Using a Cognitive Components Model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, 1996.

McKinley, D. et al. A Work-Centered Approach for Setting Passing Scores on Performance-Based Assessment. // *Evaluation & The Health Professions*, 2005, 28 (3), 349-369.

Mehrens, W. Methodological Issues in Standard Setting for Educational Exams'. // *Joint Conference on Standard Setting for Large-scale Assessment*, Proceedings: Vol.2, Ed. By L. Crocker & M. Zieky, Washington, U.S. Government Printing Office, 1994, 221-267.

Meskauskas, J. Evaluation Models for Criterion-Referenced Testing: Views Regarding Mastery and Standard-Setting. // *Review of Educational Research*, 1976, 46 (1), 133-158.

Messick, S. *Standards-Based Score Interpretation: Establishing Valid Grounds for Valid Inferences*. ETS-RR-94-57. Princeton, N.J.: Educational Testing Service, 1994.

Messick, S. Standards of Validity and the Validity of Standards in Peflormance Assessment. // *Educational Measurement: Issues and Practice*, 1995, 14 (4), 5-8.

Meulman, J. & W. Heiser. *SPSS Categories®* 14.0. Chicago, SPSS Inc., 2005.

Mills, C. & G. Melican. Estimating and Adjusting Cutoff Scores: Features of Selected Methods. // *Applied Measurement in Education*, 1988, 1 (3), 261-275.

Mislevy, R. Linking Educational Assessments: Concepts, Issues, Methods, and Prospects. Princeton, ETS, 1992.

Mitzel, H. et al. The Bookmark Procedure: Cognitive Perspectives on Standard-Setting. // *Standard-Setting: Concepts, Methods, and Perspectives*, G. J. Cizek (Ed.), Hillsdale NJ: Erlbaum, 2001, 249-282.

Morgan, D. & M. Perie. *Setting Standards in Education: Choosing the Best Method for Your Assessment and Population*. Princeton, NJ: ETS, 2004.

Muijtjens, A. et al. Using Resampling to Estimate the Precision of an Empirical Standard-Setting Method. // *Applied Measurement in Education*, 2003, 16 (3), 245-256.

Nedelsky, L. Absolute Grading Standards for Objective Tests. // *Educational and Psychological Measurement*, 1954, 14 (1), 3-19.

Nitko, A. *A model for Curriculum-Driven Criterion-Referenced and Norm-Referenced National Examinations for Certification and Selection of Students*. Paper presented at the Association for the Study of Educational Evaluation in Southern Africa's International Conference on Educational Evaluation and Assessment, Pretoria, July 1994.

Nitko, A. Distinguishing the Many Varieties of Criterion-referenced Tests. // *Review of Educational Research*, 1980, 50 (3), 461-485.

Noijons, J. & H. Kuijper. *Mapping the Dutch Foreign Language State Examinations onto the CEFR*. Report of a Cito research project commissioned by the Dutch Ministry of Education, Culture and Science, 2006, [15.01.2007], <http://toetswijzer.kennisnet.nl/html/cef/home.htm#summary>

Norcini, J. Research on Standards for Professional Licensure and Certification Examinations. // *Evaluation & the Health Professions*, 1994, 17 (2), 160-177.

Norcini, J. Setting Standards on Educational Tests. // *Medical Education*, 2003, 37, 464–469.

Norcini, J. et al. The Effect of Various Factors on Standard Setting. // *Journal of Educational Measurement*, 1988, 25(1), 57-65.

Norcini, J. & J. Shea. The Credibility and Comparability of Standards, *Applied Measurement in Education*, 1997, 10 (1), 39-59.

Pellegrino, J. et al (Eds). *Grading the Nations's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress*, Washington, National Academy Press, 1999.

Perkins, N. & E. Schisterman. The Inconsistency of ''Optimal'' Cutpoints Obtained using Two Criteria based on the Receiver Operating Characteristic Curve. // *American Journal of Epidemiology*, 2006, 163 (7), 670-675.

Plake, B. et al. Factors Influencing Intrajudge Consistency during Standard Setting. // *Educational Measurement: Issues and Practice*, 1991, 10 (2), 15-26.

Plake, B. et al. *Setting multiple performance standards using the yes/no method: An alternative item mapping procedure*. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal, 2005.

Plake, B. & G. Giraud. *Effect of a Modified Angoff Strategy for Obtaining Item Perforrnance Estimates in a Standard Setting Study*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA, 1998.

Plake, B. & J. Impara. Ability of Panelist to Estimate Item Performance for a Target Group of Candidates: An Issue in Judgmental Standard Setting. // *Educational Assessment*, 2001, 7 (2), 87-97.

Plake, B. & R. Hambleton. The Analytic Judgment Method for Setting Standards on Complex Performance Assessments. // *Standard-Setting: Concepts, Methods, and Perspectives*. G. J. Cizek (Ed.), Hillsdale NJ: Erlbaum, 2001, 283-312.

Poggio, J. & D. Glasnapp. *A Method for Setting Multi-Level Performance Standards on Objective Constructed Response Tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, 1994.

Popham, W. The Criticality of Consequences in Standard Setting: Six Lessons Learned the Hard Way by a Standard Setting Abettor. // *Proceedings of Achievement Levels Workshop – Section 7*, Boulder, 1998.

Putnam, S. et al. A Multi-Stage Dominant Profile Method for Setting Standards on Complex Performance Assessments // *Applied Measurement in Education*, 1995, 8 (1), 57-83.

Raymond, M. & J. Reid. Who Made Thee a Judge? Selecting and Training Participants for Standard Setting. // *Standard-setting: Concepts, methods, and perspectives*. Ed. by G. Cizek, Hillsdale, Erlbaum, 2001, 119-157.

Reckase, M. *Analysis of Methods for Collecting Test-based Judgments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, 1998.

Reckase, M. A Survey and Evaluation of Recently Developed Procedures for Setting Standards on Educational Tests. // *Student performance Standards on the National Assessment of Educational progress: Affirmations and Improvement*. Ed. By M. Bourque & Sh. Byrd, Washington, NAEP, 2000-a, 41 – 70.

Reckase, M. *The ACT/NAGB Standard Setting Process: How "Modified" Does It Have To Be before It Is No Longer a Modified-Angoff Process?* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, 2000-б.

Reckase, M. A Conceptual Framework for a Psychometric Theory for Standard Setting with Examples of Its Use for Evaluating the Functioning of Two Standard Setting Methods. // *Educational Measurement: Issues and Practice*, 2006, 25 (2), 4-18.

Reid, J. Training Judges to Generate Standard-Setting Data. *Educational Measurement: Issues and Practice*, 1991, 10 (2), 11-14.

Reilly, R. et al. Comparison of Direct and Indirect Methods For Setting Minimum Passing Scores. // *Applied Psychological Measurement*, 1984, 8 (4), 421-429.

Rodgers, J. The Bootstrap, the Jackknife, and the Randomization Test: A Sampling Taxonomy. // *Multivariate Behavioral Research*, 1999, 34 (4), 441-456.

Sadesky, Gr. & M. Gushta. *Standard Setting Using the Attribute Hierarchy Model*, Paper presented at the annual meeting of the National Council on Measurement in Education, 2004.

Sadler, D. Interpretations of criteria-based assessment and grading in higher education. // *Assessment & Evaluation in Higher Education*, 2005, 30 (2), 175-194.

Schulz, E. et al. A Rationale for Defining Achievement Levels Using IRT-Estimated Domain Scores. // *Applied Psychological Measurement*, 1999, 23 (4), 347–362.

Schultz, E., & H. Mitzel, *The Mapmark Standard Setting Method*. Paper presented to the National Assessment Governing Board for the National Assessment of Educational Progress, 2005.

Shepard, L. Standard Setting Issues and Methods. // *Applied Psychological Measurement*, 1980-б, 4 (4), 447-467.

Shepard, L. Technical Issues in Minimum Competency Testing. // *Review of Research in Education*, Vol. 8, Ed. By D. Berliner, 1980-b, Itasca, F. E. Peacock Publishers, 30-82.

Sireci, S. Standard Setting Using Cluster Analysis. // *Standard-Setting: Concepts, Methods, and Perspectives*. Ed. by G. J. Cizek, Hillsdale NJ: Erlbaum, 2001, 339-354.

Smith, R. & J. Smith. Differential Use of Item Information by Judges Using Angoff and Nedelsky Procedures. // *Journal of Educational Measurement*, 1988, 25 (4), 259-274.

Stoyanova, F. Fuzzy Set Theory Application in Criterion - Referenced Testing. // *Language Testing: New Openings*, Ed. by A. Hunta et al, University of Jyvaskyla, 1993, 103-111.

Subkoviak, M. Estimating Reliability from a Single Administration of a Criterion-referenced Test. // *Journal of Educationa1 Measurement*, 1976, 13 (4), 265-275.

Subkoviak, M. Decision-consistency Approaches. // *Criterion-referenced Measurement: The State of the Art*. Ed. by R. Berk, Baltimore, The Johns Hopkins University Press, 1980.

Subkoviak, M. A Practitioner's Guide to Computation and Interpretation of Reliability Indices for Mastery Tests. // *Journal of Educational Measurement*, 1988, 25 (1), 47-55.

Swaminathan, H. et al. Reliability of Criterion-referenced tests: A Decision-theoretic Formulation. // *Journal of Educational Measurement*, 1974. 11 (4), 263-268.

Swets, J. et al. Psychological science can improve diagnostic decisions. // *Psychological Science in the Public Interest*, 2000, 1(1), 1–26.

Takane, Y. Applications of multidimensional scaling in psychometrics. // *Handbook of Statistics: Pyschometrics*, V. 26, Ed. by C. Rao, & S. Sinharay, Amsterdam, Elsevier, 2006, 359-400.

Thompson, Br. The Pivotal Role of Replication in Psychological Research: Empirically Evaluating the Replicability of Sample Results. // *Journal of Personality*, 1994, 62(2), 157-176.

van der Linden, W. A Latent Trait Method for Determining Intrajudge Inconsistency in the Angoff and Nedelsky Techniques of Standard Setting. // *Journal of Educational Measurement*, 1982, 19 (4), 295 – 308.

van der Linden, W. *A Conceptual Analysis of Standard Setting in Large-Scale Assessments*. Research Report 94-3, Twente University, Faculty of Educational Science and Technology, 1994.

van der Schoot, F. *IRT-based Method for Standard Setting in a Three-Stage Procedure*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, 2002.

Verhelst, N. et al. One-Parameter Logistic Model: OPLM. Arnhem, Cito, 1995.

Verhelst, N. & F. Kaftandjieva. *A Rational Method to Determine Cutoff Scores*. Research Report 99–0, Enschede, The Netherlands: University of Twente, Faculty of Educational Science and Technology, Department of Educational Measurement and Data Analysis, 1999.

Wan, L. et al. *Estimating Classification Consistency for Complex Assessments*. CASMA Research Report 22. Iowa, Center for Advanced Studies in Measurement and Assessment, 2007.

Wang, N. Use of the Rasch IRT Model in Standard Setting: An Item-Mapping Method // *Journal of Educational Measurement*, 2003, 40 (3), 231-253.

White, A. *Result Generalizability and Detection of Discrepant Data Points: Illustrating the Jackknife Method*. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Dallas, 2000.

Wiliam, D. *Construct-referenced Assessment of Authentic tasks: Alternatives to Norms and Criteria*. Paper presented at the 24th Annual Conference of the International Association for Educational Assessment – Testing and Evaluation: Confronting the Challenges of Rapid Social Change, Barbados, May, 1998.

Wright, B. & N. Masters. *Rating Scale Analysis*. Chicago: MESA Press, 1982.

Wright, B. Reliability and Separation. // *Rasch Measurement Transactions*, 1996, 9:4, 472, [8.12.1999], <http://209.41.24.153/rmt/rmt94.htm>

Young, F. & D. Harris. Multidimensional scaling. // *SPSS for Windows: Professional statistics*. Ed. by M. Noursis, Chicago, SPSS, Inc., 1993, 157-222.

Yu, Ch. Resampling methods: Concepts, Applications, and Justification. // *Practical Assessment, Research & Evaluation*, 2003, 8 (19).

Zieky, M. A Historical Perspective on Setting Standards. // *Joint Conference on Standard Setting for Large-Scale Assessments*. Proceedings: Vol.2, Ed. By L. Crocker & M. Zieky, Washington, U.S. Government Printing Office, 1994, 1-38.

Zieky, M. So Much Has Changed: How the Setting of Cutscores Has Evolved Since 1980. // *Standard-setting: Concepts, methods, and perspectives*. Ed. By G. Cizek, Hillsdale NJ: Erlbaum, 2001, 19-52

Appendices

Appendix 1

List of the existing methods for setting cut scores

№	Method	Source	Year	Judgment			Limitations			Cut scores	
				Focus	Stages	Emp. data	Format	Scoring	Model	Emp. data	Stat. meth.
1	Nedelsky	Nedelsky, 1954	1954	test items	1	no	m. c. resp.	dich.	no	no	descr. stat.
2	Angoff (original)	Angoff, 1971	1971	test items	1	no	no	dich.	no	no	descr. stat.
3	Angoff (footnote)	Angoff, 1971	1971	test items	1	no	no	dich.	no	no	descr. stat.
4	Ebel	Ebel, 1972	1972	test items	2	no	no	dich.	no	no	descr. stat.
5	Modified Angoff	Plake, B. & G. Giraud, 1998	1976	test items	1	no	no	dich.	no	no	descr. stat.
6	Jaeger	Jaeger, 1982	1982	test items	3	yes	no	dich.	no	yes	descr. stat.
7	Borderline Group	Livingston & Zieky, 1982	1982	test takers	1	no	no	no	no	yes	descr. stat.
8	Contrasting Groups	Livingston & Zieky, 1982	1982	test takers	1	no	no	no	no	yes	descr. stat.+
9	Up-and-down	Livingston & Zieky, 1982	1982	item resp.	1	yes	ext. resp.	no	no	yes	descr. stat.
10	Reference group	Livingston & Zieky, 1982	1982	test takers	1	no	no	no	no	yes	descr. stat.
11	Hofstee	Mills & Melican, 1988	1983	test score	1	no	no	no	no	yes	descr. stat.+
12	Nedelsky (modified)	Reilly et al, 1984	1984	test items	1	no	m. c. resp.	dich.	no	no	descr. stat.+ prob. theor.
13	Iterative Angoff	Berk, 1986	1984	test items	3	yes	no	dich.	no	no	descr. stat.
14	Beuk	Beuk, 1984	1984	test score	1	no	no	no	no	yes	descr. stat.+
15	Weighted Objectives	Cangelosi, 1984	1984	test items+	1	no	no	no	no	no	descr. stat.
16	de Gruijter	de Gruijter, 1985	1985	test score	1	no	no	no	no	yes	descr. stat.+
17	Direct standard setting	Jones et al, 1988	1988	test items	>1	no	no	no	no	no	descr. stat.
18	Combined Judgment-Empirical	Livingston, 1991	1991	test items+	2	no	no	dich.	IRT	yes	descr. stat.+ IRT
19	Fuzzy sets	Stoyanova, 1993	1991	test items	1	no	no	no	no	no	descr. stat.+ fuzzy sets
20	Body of work	Kingston et al., 2001	1992	item resp.	2	yes	no	no	no	yes	descr. stat.+ log. regr.
21	Paper Selection	Loomis & Bourque, 2001	1992	item resp.	3	no	ext. resp.	polit.	no	yes	descr. stat.

№	Method	Source	Year	Judgment			Limitations			Cut scores	
				Focus	Stages	Emp. data	Format	Scoring	Model	Emp. data	Stat. meth.
22	Objective standard setting	Wright & Grosse, 1993	1993	test items+	2	yes	no	dich.	IRT	yes	descr. stat.+
23	Angoff – Derivatives	Loomis & Bourque, 2001	1994	test items	1	no	no	no	no	no	descr. stat.
24	Extended Angoff	Hambleton & Plake, 1995	1994	test items+	4	no	ext. resp.	polit.	no	no	descr. stat.
25	Estimated Mean	Loomis & Bourque, 2001	1994	test items	3	no	no	no	no	no	descr. stat.
26	Item Score Distribution	Poggio & Glasnapp, 1994	1994	test score	1	no	no	no	no	no	descr. stat.
27	Percent correct	Loomis & Bourque, 2001	1994	test items	1	no	no	no	no	no	descr. stat.
28	Reckase	Loomis & Bourque, 2001	1994	test items	3	yes	no	no	IRT	yes	descr. stat.+ IRT
29	Mean estimation	Loomis & Bourque, 2001	1994	test items	1	no	no	no	no	no	descr. stat.
30	Proportional correct	Loomis & Bourque, 2001	1994	test items	1	no	no	no	no	no	descr. stat.
31	Judgmental Policy Capturing	Jaeger, 1995	1995	profile	2	no	ext. resp.	polit.	no	yes	descr. stat.+ mult. regr.
32	Dominant Profile	Putnam et al., 1995	1995	profile	3	no	ext. resp.	polit.	no	no	descr. stat.
33	Cluster	Sireci, 2001	1995	test items	1	no	no	no	no	yes	descr. stat.+ clust. an.
34	Bookmark	Mitzel et al., 2001	1996	test items+	3	yes	no	no	IRT	yes	descr. stat.+ IRT
35	Integrated Judgment	Jaeger & Mills, 2001	1996	item resp.	1	yes	no	no	no	yes	descr. stat.+ lin. regr.
36	Cognitive Components	McGinty & Neel, 1996	1996	test items+	2	no	no	no	no	no	descr. stat.+
37	Advanced impact method	Impara & Plake, 2000	1996	test score	1	no	no	no	no	no	descr. stat.
38	Adjusted Angoff	Taube, 1997	1997	test items	4	no	no	dich.	IRT	yes	descr. stat.+ IRT
39	Angoff 'Yes/No'	Impara & Plake, 1997	1997	test items	2	yes	no	dich.	no	yes	descr. stat.
40	Item Score String Estimation	Loomis & Bourque, 2001	1997	test items	2	yes	no	dich.	no	no	descr. stat.
41	Grid	Loomis & Bourque, 2001	1997	profile	1	no	no	polit.	no	no	descr. stat.
42	Booklet Classification	Loomis & Bourque, 2001	1998	item resp.	1	yes	no	no	no	yes	descr. stat.+ nonl. regr.
43	Multistage Aggregation	De Champlain et al, 1998	1998	test items+	4	yes	no	no	no	yes	descr. stat.+ log. regr.
44	IDEA	Chang et al., 2004	1999	test items	1	no	m. c. resp.	dich.	no	no	descr. stat.

№	Method	Source	Year	Judgment			Limitations			Cut scores	
				Focus	Stages	Emp. data	Format	Scoring	Model	Emp. data	Stat. meth.
45	Item Mastery	Verhelst & Kaftandjieva, 1999	1999	test items	1	no	no	no	IRT	yes	descr. stat.+ math. opt.
46	Item-Domain Levels	Schulz et al., 1999	1999	test items+	1	no	no	dich.	IRT	yes	descr. stat.+ IRT
47	Generalized Examinee-Centered	Cohen, Kane & Crooks, 1999	1999	item resp.	1	yes	no	no	no	yes	descr. stat.+ nonl. regr.
48	Hierarchical IRT	Janssen et al., 2000	2000	test items	1	no	no	dich.	IRT	yes	descr. stat.+ IRT
49	Compound Cumulative	Kaftandjieva & Takala, 2002	2001	test items	1	no	no	no	no	yes	descr. stat.+
50	Basket procedure	Council of Europe, 2003	2001	test items	1	no	no	no	no	no	descr. stat.
51	Analytical Judgment	Plake & Hambleton, 2001	2001	item resp.	2	no	no	no	no	yes	descr. stat.+
52	Generalized Holistic	Cizek & Bunch, 2007	2001	item resp.	2	no	ext. resp.	polit.	no	no	descr. stat.
53	Multistage IRT	van der Schoot, 2002	2002	test items+	3	yes	no	dich.	IRT	yes	descr. stat.+
54	Item Descriptor Matching	Cizek & Bunch, 2007	2002	test items+	3	yes	no	no	IRT	yes	descr. stat.+
55	Item mapping	Wang, 2003	2003	test items+	2	yes	no	dich.	IRT	yes	descr. stat.+ IRT
56	Performance Profile	Morgan & Perie, 2004	2003	profile	3	yes	ext. resp.	polit.	no	yes	descr. stat.
57	Item Signature Method	Hattie & Brown, 2003	2003	test items	2	no	no	no	no	no	descr. stat.
58	Alternative Item Mapping	Plake et al, 2005	2003	test items	2	yes	no	no	no	yes	descr. stat.
59	Direct consensus	Cizek & Bunch, 2007	2004	test items	3	no	no	no	no	no	descr. stat.
60	Attribute Hierarchy	Sadesky & Gushta, 2004	2004	test items	1	no	no	no	IRT	yes	descr. stat.+
61	Mapmark	Schultz & Mitzel, 2005	2005	test items+	4	yes	no	no	IRT	yes	descr. stat.+ IRT
62	Work-centered	McKinley, D. et al., 2005	2005	profile	1	yes	no	no	no	no	descr. stat.+ nonl. regr.

Appendix 2

Illustrative example: test items and judgment

Item №	Discr. index[1]		Difficulty[2]			Model matching[3]		Judg-ment[4]		Dichotomization[5]					
										Level A2		Level B1		Level B2	
	D_R	D_S	%		Z	P_R	P_S	E1	E2	$A2_1$	$A2_2$	$B1_1$	$B1_2$	B21	$B2_2$
			R	*S*											
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
1	+0.49	+0.52	29	28	+1.29	0.200	0.021	4	4	0	0	0	0	1	1
2	+0.54	+0.56	51	50	-0.06	0.390	0.137	3	3	0	0	1	1	1	1
3	+0.64	+0.57	64	64	-0.78	0.070	0.595	2	2	1	1	1	1	1	1
4	+0.68	+0.58	56	57	-0.36	0.021	0.910	2	2	1	1	1	1	1	1
5	+0.54	+0.58	53	53	-0.15	0.742	0.823	2	3	1	0	1	1	1	1
6	+0.68	+0.58	48	48	+0.15	0.050	0.260	2	3	1	0	1	1	1	1
7	+0.69	+0.57	56	56	-0.31	0.033	0.625	2	3	1	0	1	1	1	1
8	+0.56	+0.56	63	63	-0.73	0.425	0.383	3	2	0	1	1	1	1	1
9	+0.43	+0.57	54	54	-0.22	0.062	0.005	3	3	0	0	1	1	1	1
10	+0.33	+0.39	88	87	-2.61	0.893	0.018	3	2	0	1	1	1	1	1
11	+0.49	+0.56	56	57	-0.36	0.699	0.398	5	2	0	1	0	1	0	1
12	+0.44	+0.55	66	66	-0.95	0.263	0.218	2	2	1	1	1	1	1	1
13	+0.49	+0.58	48	48	+0.12	0.129	0.940	3	3	0	0	1	1	1	1
14	+0.63	+0.52	25	25	+1.60	0.175	0.688	5	4	0	0	0	0	0	1
15	+0.49	+0.47	16	17	+2.34	0.788	0.091	3	5	0	0	1	0	1	0
16	+0.54	+0.49	21	22	+1.88	0.402	0.720	3	4	0	0	1	0	1	1
17	+0.70	+0.57	49	50	+0.05	0.075	0.814	2	3	1	0	1	1	1	1
18	+0.68	+0.57	38	37	+0.74	0.044	0.951	4	3	0	0	0	1	1	1
19	+0.50	+0.54	33	32	+1.05	0.281	0.516	3	3	0	0	1	1	1	1
20	+0.52	+0.57	46	46	+0.24	0.428	0.922	4	3	0	0	0	1	1	1
21	+0.58	+0.54	66	66	-0.92	0.154	0.451	3	2	0	1	1	1	1	1
22	+0.51	+0.55	29	29	+1.29	0.064	0.686	3	4	0	0	1	0	1	1
23	+0.52	+0.59	52	52	-0.13	0.744	0.287	3	3	0	0	1	1	1	1
24	+0.51	+0.58	55	55	-0.27	0.590	0.527	4	3	0	0	0	1	1	1
25	+0.41	+0.48	20	19	+1.95	0.608	0.111	3	5	0	0	1	0	1	0
26	+0.40	+0.42	87	86	-2.44	0.369	0.816	2	2	1	1	1	1	1	1
27	+0.34	+0.43	86	85	-2.40	0.651	0.397	2	2	1	1	1	1	1	1

22 The indices ***R*** and ***S*** in both columns for the discrimination index of the items correspond to the two samples: ***R*** for the real data (n = 250) and ***S*** for simulated data (n = 5000)

23 In the first two colums (***R*** and ***S***), the item difficulty in the real (***R***) and simulated (***S***) data is expressed in %, and in the third column (***Z***) – in Z-scale, adjusted against the item difficulty (Z_{mean} = 0; SD_Z = 1).

24 ***P*** – the probability for match between the concrete data (real – ***R*** or simulated – ***S***) and the chosen theoretical model for the scaling of the item difficulty and the ability for reading comprehension of the examinees (one-parameter Rasch model).

25 In the judgment the following coding of the data was used: 1 – A1; 2 – A2; 3 – B1; 4 – B2; 5 – C1; 6 – C2.

26 In the dichotomization, code '**1**' means that if given examinee is below this level of competence (A2 or B1), he or she has to be able to answer **correctly** the corresponding item according the particular judge (E1 or E2). Code '**0**' means that, acording to the expert, the examinees which are at that level of competence could not answer the respective item correctly.

Appendix 3

Illustrative example: Table for transformation of the raw test score into Z-scale and cut scores

Raw test score (CSEM)[1]	Z-scale (CSEM)	Frequencies n =		Level of competence2											
				M1		M2		M3		M4		M5		M6	
		250	5000	E1	E2	E1	E2	E1	E2	E1	E1	E1	E2	E1	E2
0	-4.86 (1.866)	0	28	≤A2	≤A2	≤A2	≤A2	≤A2	≤A2	≤A2	≤A2	≤A2	≤A2	≤A2	≤A2
1 (0,87)	-3.63 (0.979)	1	69												
2 (1,20)	-2.98 (0.766)	1	102												
3 (1,44)	-2.52 (0.662)	7	145												
4 (1,63)	-2.14 (0.598)	10	159												
5 (1,78)	-1.82 (0.553)	8	197												
6 (1,91)	-1.55 (0.521)	20	199												
7 (2,01)	-1.30 (0.496)	12	215												
8 (2,09)	-1.07 (0.477)	12	233				B1		B1						
9 (2,16)	-0.85 (0.463)	15	220												
10 (2,21)	-0.65 (0.453)	14	228	B1	B1	B1		B1		B1	B1	B1	B1	B1	B1
11 (2,25)	-0.46 (0.445)	10	249												
12 (2,28)	-0.27 (0.441)	7	234												
13 (2,29)	-0.08 (0.438)	12	226												
14 (2,29)	+0.11 (0.438)	13	240												
15 (2,28)	+0.30 (0.441)	8	250												
16 (2,25)	+0.49 (0.445)	8	216											B1	
17 (2,21)	+0.69 (0.452)	12	236											B2	
18 (2,16)	+0.89 (0.462)	12	232											C1≥	
19 (2,09)	+1.10 (0.475)	7	209								B2	B2			
20 (2,01)	+1.33 (0.491)	14	202			B2≥		B2≥							
21 (1,91)	+1.57 (0.513)	7	206				B2		B2	B2		C1≥	B2		B2
22 (1,78)	+1.84 (0.542)	10	157	B2	B2						B2			C1≥	
23 (1,63)	+2.14 (0.583)	10	169										C1≥		
24 (1,44)	+2.49 (0.644)	6	129							C1≥					C1≥
25 (1,20)	+2.92 (0.745)	8	112				C1≥		C1≥		C1≥				
26 (0,87)	+3.53 (0.957)	5	89	C1≥	C1≥	C1≥		C1≥							
27	+4.73 (1.833)	1	49												

M1 Basket procedure
M2 Cumulative Compound method
M3 Cumulative Cluster method
M4 ROC-curve method
M5 Item Mastery method
M6 Level Characteristic Curves method

27 For each concrete value of the test score (number of correctly answered test items or in Z-scale) the conditional standard error is provided in brackets.

28 The standards of the Common European Framework of Reference for Languages are used to designate the levels of competence.

Appendix 4

Common European Framework of Reference for Languages: Descriptors for reading

A1	• Can understand familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. (p. 24) • Can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues. (p. 26) • Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required. (p. 69) • Can understand short, simple messages on postcards. (p. 69) • Can recognise familiar names, words and very basic phrases on simple notices in the most common everyday situations. (p. 70) • Can get an idea of the content of simpler informational material and short simple descriptions, especially if there is visual support. (p. 70) • Can follow short, simple written directions (e.g. to go from X to Y). (p. 71)
A2	• Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). (p. 24) • Can read very short, simple texts. (p. 26) • Can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus and timetables. (p. 26) • Can understand short simple personal letters. (p. 26) • Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items. (p. 69) • Can understand basic types of standard routine letters and faxes (enquiries, orders, letters of confirmation etc.) on familiar topics. (p. 69) • Can locate specific information in lists and isolate the information required (e.g. use the 'Yellow Pages' to find a service or tradesman). (p. 70) • Can understand everyday signs and notices: in public places, such as streets, restaurants, railway stations; in workplaces, such as directions, instructions, hazard warnings. (p. 70) • Can identify specific information in simpler written material he/she encounters such as letters, brochures and short newspaper articles describing events. (p. 70) • Can understand simple instructions on equipment encountered in everyday life – such as a public telephone. (p. 71) • Can understand regulations, for example safety, when expressed in simple language. (p. 71) • Can use an idea of the overall meaning of short texts and utterances on everyday topics of a concrete type to derive the probable meaning of unknown words from the context. (p. 72)
B1	• Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. (p. 24) • Can understand texts that consist mainly of high frequency everyday or job-related language. (p. 26) • Can understand the description of events, feelings and wishes in personal letters. (p. 26) • Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension. (p. 69) • Can understand the description of events, feelings and wishes in personal letters well enough to correspond regularly with a pen friend. (p. 69) • Can find and understand relevant information in everyday material, such as letters, brochures and short official documents. (p. 70) • Can scan longer texts in order to locate desired information, and gather information from different parts of a text, or from different texts in order to fulfil a specific task. (p. 70) • Can recognise significant points in straightforward newspaper articles on familiar subjects. (p. 70) • Can identify the main conclusions in clearly signalled argumentative texts. (p. 70) • Can recognise the line of argument in the treatment of the issue presented, though not necessarily in detail. (p. 70) • Can understand clearly written, straightforward instructions for a piece of equipment. (p. 70) • Can identify unfamiliar words from the context on topics related to his/her field and interests. (p. 72) • Can extrapolate the meaning of occasional unknown words from the context and deduce sentence meaning provided the topic discussed is familiar. (p. 72)

B2	• Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. (p. 24) • Can read articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints. (p. 27) • Can understand contemporary literary prose. (p. 27) • Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. (p. 69) • Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms. (p. 69) • Can read correspondence relating to his/her field of interest and readily grasp the essential meaning. (p. 69) • Can scan quickly through long and complex texts, locating relevant details. (p. 70) • Can quickly identify the content and relevance of news items, articles and reports on a wide range of professional topics, deciding whether closer study is worthwhile. (p. 70) • Can understand articles and reports concerned with contemporary problems in which the writers adopt particular stances or viewpoints. (p. 70) • Can obtain information, ideas and opinions from highly specialised sources within his/her field. (p. 70) • Can understand specialised articles outside his/her field, provided he/she can use a dictionary occasionally to confirm his/her interpretation of terminology. (p. 70) • Can understand lengthy, complex instructions in his field, including details on conditions and warnings, provided he/she can reread difficult sections. (p. 70)
C1	• Can understand a wide range of demanding, longer texts, and recognise implicit meaning. (p. 24) • Can understand long and complex factual and literary texts, appreciating distinctions of style. (p. 27) • Can understand specialised articles and longer technical instructions, even when they do not relate to my field. (p. 27) • Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of speciality, provided he/she can reread difficult sections. (p. 69) • Can understand any correspondence given the occasional use of a dictionary. (p. 69) • Can understand in detail a wide range of lengthy, complex texts likely to be encountered in social, professional or academic life, identifying finer points of detail including attitudes and implied as well as stated opinions. (p. 70) • Can understand in detail lengthy, complex instructions on a new machine or procedure, whether or not the instructions relate to his/her own area of speciality, provided he/she can reread difficult sections. (p. 70) • Is skilled at using contextual, grammatical and lexical cues to infer attitude, mood and intentions and anticipate what will come next. (p. 72)
C2	• Can understand with ease virtually everything read. (p. 24) • Can read with ease virtually all forms of the written language, Including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works. (p. 27) • Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings. (p. 69) • Can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning. (p. 69)

Appendix 5

ROC-curve method: Indices of sensitivity and specificity and optimization criteria

№	Cut score	E1 – Level A2				E1 – Level B1				E1 – Level B2			
		S[1]	*1 – P*[2]	*J*[3]	D_{01}[4]	*S*	*1 – P*	*J*	D_{01}	*S*	*1 – P*	*J*	D_{01}
1	-3.61	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
2	-2.52	0.00	0.06	-0.06	1.00	0.05	0.00	0.05	0.95	0.04	0.00	0.04	0.96
3	-2.42	0.11	0.06	0.06	0.89	0.10	0.00	0.10	0.91	0.08	0.00	0.08	0.92
4	-1.68	0.22	0.06	0.17	0.78	0.14	0.00	0.14	0.86	0.12	0.00	0.12	0.88
5	-0.93	0.33	0.06	0.28	0.67	0.19	0.00	0.19	0.81	0.16	0.00	0.16	0.84
6	-0.85	0.33	0.11	0.22	0.68	0.24	0.00	0.24	0.76	0.20	0.00	0.20	0.80
7	-0.75	0.44	0.11	0.33	0.57	0.29	0.00	0.29	0.71	0.24	0.00	0.24	0.76
8	-0.55	0.44	0.17	0.28	0.58	0.33	0.00	0.33	0.67	0.28	0.00	0.28	0.72
9	-0.34	0.56	0.22	0.33	0.50	0.38	0.17	0.21	0.64	0.32	0.50	-0.18	0.84
10	-0.29	0.67	0.22	0.45	0.40	0.43	0.17	0.26	0.59	0.36	0.50	-0.14	0.81
11	-0.25	0.67	0.28	0.39	0.43	0.43	0.33	0.10	0.66	0.40	0.50	-0.10	0.78
12	-0.19	0.67	0.33	0.33	0.47	0.48	0.33	0.14	0.62	0.44	0.50	-0.06	0.75
13	-0.14	0.78	0.33	0.45	0.40	0.52	0.33	0.19	0.58	0.48	0.50	-0.02	0.72
14	-0.10	0.78	0.39	0.39	0.45	0.57	0.33	0.24	0.54	0.52	0.50	0.02	0.69
15	-0.01	0.78	0.44	0.34	0.50	0.62	0.33	0.29	0.51	0.56	0.50	0.06	0.67
16	0.09	0.89	0.44	0.45	0.46	0.67	0.33	0.34	0.47	0.60	0.50	0.10	0.64
17	0.13	0.88	0.50	0.39	0.51	0.71	0.33	0.38	0.44	0.64	0.50	0.14	0.61
18	0.19	1.00	0.50	0.50	0.50	0.76	0.33	0.43	0.41	0.68	0.50	0.18	0.59
19	0.49	1.00	0.56	0.44	0.56	0.76	0.50	0.26	0.55	0.72	0.50	0.22	0.57
20	0.89	1.00	0.61	0.39	0.61	0.76	0.67	0.10	0.71	0.76	0.50	0.26	0.55
21	1.17	1.00	0.67	0.33	0.67	0.81	0.67	0.14	0.69	0.80	0.50	0.30	0.54
22	1.44	1.00	0.78	0.22	0.78	0.86	0.83	0.02	0.85	0.88	0.50	0.38	0.51
23	1.74	1.00	0.83	0.17	0.83	0.86	1.00	-0.14	1.01	0.88	1.00	-0.12	1.01
24	1.92	1.00	0.89	0.11	0.89	0.91	1.00	-0.10	1.01	0.92	1.00	-0.08	1.00
25	2.15	1.00	0.94	0.06	0.94	0.95	1.00	-0.05	1.00	0.96	1.00	-0.04	1.00
26	3.34	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00

29 ***S*** – sensitivity index

30 ***P*** – specificity index

31 ***J*** – Youden index

32 D_{01} – distance to a point (0; 1)

Appendix 6A

Cut scores for the different judges and methods (Raw test score)

Test	Judge	Cut score *X*					Cut score *Y*			
		M1	M2	M3	M4		M1	M2	M3	M4
T1	E_{101}	14	24	22	24		39	33	32	28
	E_{102}	18	26	25	22		42	33	36	34
	E_{103}	19	26	22	26		37	33	36	32
	E_{104}	19	26	26	24		44	36	36	28
	E_{105}	10	18	18	28		33	32	32	28
	E_{106}	12	24	22	18		40	32	32	18
	E_{107}	17	24	24	21		37	33	33	24
T2	E_{201}	31	28	27	28		44	39	39	30
	E_{202}	18	24	26	16		42	31	36	36
	E_{203}	26	26	26	21		48	39	36	36
	E_{204}	5	21	21	13		34	36	30	35
	E_{205}	26	28	26	29		49	39	43	47
	E_{206}	29	26	27	28		45	36	36	30
	E_{207}	28	26	26	21		42	36	36	29
	E_{208}	23	25	26	14		44	36	36	29
	E_{209}	25	28	28	26		50	42	43	47
	E_{210}	29	26	26	27		40	36	36	29
T3	E_{301}	19	24	24	27		36	27	27	21
	E_{302}	11	21	21	19		34	27	28	14
	E_{303}	15	21	21	20		33	28	29	21
	E_{304}	26	27	24	13		39		24	
	E_{305}	7	20	20	18		31	27	27	19
	E_{306}	33	27	27	28		39		34	
	E_{307}	9	19	20	14		25	24	27	29
	E_{308}	21	24	24	27		39		31	
	E_{309}	10	19	21	18		34	30	28	34
	E_{310}	21	24	24	28		35	27	27	21
	E_{311}	11	20	24	21		30	27	27	19
	E_{312}	12	19	21	27		24	27	27	27
	E_{313}	7	24	24	29		33	27	28	29

M1 Basket procedure
M2 Cumulative Compound method
M3 Cumulative Cluster method
M4 ROC-curve method

Appendix 6B

Cut scores for the different judges and methods (Z-scale)

Test	Judge	Cut score *X*							Cut score *Y*					
		M1	**M2**	**M3**	**M4**	**M5**	**M6**		**M1**	**M2**	**M3**	**M4**	**M5**	**M6**
T1	E_{101}	-0.28	-0.59	-0.59	-0.07	-0.53	+0.04		+0.66	+0.33	+0.33	-0.02	+0.42	+0.34
	E_{102}	-0.13	-0.39	-0.39	-0.07	-0.35	+0.15		+0.83	+0.37	+0.33	+0.30	+0.52	+0.33
	E_{103}	-0.12	-0.39	-0.45	-0.12	-0.33	+0.15		+0.53	+0.37	+0.33	+0.33	+0.33	+0.19
	E_{104}	-0.06	-0.39	-0.39	-0.02	-0.24	+0.10		+0.86	+0.50	+0.50	-0.02	+0.59	+0.53
	E_{105}	-0.50	-0.84	-0.84	-0.07	-0.54	-0.08		+0.33	+0.30	+0.33	-0.02	+0.27	+0.38
	E_{106}	-0.42	-0.84	-0.84	-0.07	-0.32	+0.01		+0.62	+0.33	+0.33	-0.34	+0.39	+0.25
	E_{107}	-0.18	-0.45	-0.45	-0.07	-0.29	+0.00		+0.55	+0.33	+0.33	-0.02	+0.39	+0.33
T2	E_{201}	+0.28	+0.18	+0.11	+0.16	+0.27	+0.16		+0.96	+0.65	+0.59	+0.44	+0.53	+0.53
	E_{202}	-0.10	+0.05	+0.16	-0.16	-0.03	+0.04		+0.73	+0.30	+0.30	+0.52	+0.40	+0.14
	E_{203}	0.15	+0.11	+0.05	+0.27	+0.09	+0.10		+0.29	+0.65	+0.47	+0.52	+0.69	+0.66
	E_{204}	-0.69	-0.04	-0.10	+0.30	-0.47	+0.02		+0.39	+0.52	+0.16	+0.44	+0.21	+0.06
	E_{205}	+0.06	+0.18	+0.11	+0.07	+0.17	+0.27		+1.53	+0.65	+0.79	+1.28	+0.86	+0.86
	E_{206}	+0.25	+0.11	+0.11	+0.16	+0.26	+0.26		+1.01	+0.52	+0.59	+0.16	+0.58	+0.49
	E_{207}	+0.20	+0.11	+0.05	+0.44	+0.19	+0.09		+0.79	+0.52	+0.56	+0.16	+0.52	+0.47
	E_{208}	+0.05	+0.07	+0.05	+0.03	+0.06	+0.09		+0.89	+0.52	+0.56	+0.16	+0.57	+0.42
	E_{209}	+0.01	+0.18	+0.11	+0.07	+0.18	+0.25		+1.79	+0.85	+0.79	+1.53	+0.93	+0.93
	E_{210}	+0.17	+0.11	+0.05	+0.16	+0.26	+0.14		+0.65	+0.52	+0.56	+0.16	+0.45	+0.35
T3	E_{301}	+0.06	+0.01	+0.01	+0.41	-0.13	-0.02		+0.96	+0.32	+0.32	+0.18	+0.70	+0.12
	E_{302}	-0.29	-0.17	-0.11	-0.42	-0.15	+0.01		+0.66	+0.32	+0.32	-0.17	+0.48	+0.19
	E_{303}	-0.09	-0.14	-0.14	-0.24	-0.27	+0.00		+0.61	+0.32	+0.32	+0.18	+0.49	+0.31
	E_{304}	+0.21	+0.21	+0.21	-0.42	0.23	+0.03		+2.29		+0.32			
	E_{305}	-0.45	-0.21	-0.26	+0.21	-0.39	-0.06		+0.44	+0.21	+0.21	+0.18	+0.34	+0.15
	E_{306}	+0.69	+0.21	+0.32	+0.41	+0.32	+0.29		2.29		+0.87			
	E_{307}	-0.33	-0.05	-0.05	-0.05	-0.31	-0.03		+0.28	+0.18	+0.18	+0.46	+0.00	+0.10
	E_{308}	+0.05	+0.10	+0.14	+0.32	+0.03	+0.09		+2.29		+0.79			
	E_{309}	-0.33	-0.24	-0.28	+0.01	-0.29	-0.02		+0.79	+0.58	+0.32	+0.63	+0.56	+0.66
	E_{310}	+0.08	-0.17	-0.14	+0.18	-0.02	-0.04		+0.83	+0.32	+0.32	+0.18	+0.55	+0.19
	E_{311}	-0.24	-0.17	-0.21	+0.18	-0.24	-0.03		+0.39	+0.32	+0.32	+0.18	+0.22	+0.23
	E_{312}	-0.19	-0.45	-0.14	+0.25	-0.35	+0.01		+0.14	+0.18	+0.18	+0.32	+0.10	+0.08
	E_{313}	-0.43	0.10	-0.42	+0.46	-0.33	-0.04		+0.66	+0.18	+0.32	+0.46	+0.30	-0.06

M1 Basket procedure
M2 Cumulative Compound method
M3 Cumulative Cluster method
M4 ROC-curve method
M5 Item Mastery method
M6 Level Characteristic Curves method

Appendix 6C

Cut scores – descriptive statistics

Data	Cut scores and scales		Method	Test 1			Test 2			Test 3		
				C	SE_c	SD_c	*C*	SE_c	SD_c	*C*	SE_c	SD_c
Original data	Cut score *X*	Raw test score	M1	15.57	1.36	3.60	24.00	2.41	7.62	15.54	2.20	7.95
			M2	24.00	1.07	2.83	25.80	.68	2.15	22.23	.81	2.92
			M3	22.71	.99	2.63	25.90	.59	1.85	22.69	.59	2.14
			M4	23.29	1.25	3.30	22.30	1.96	6.18	22.23	1.57	5.67
		Z-scale	M1	-0.24	0.06	0.16	+0.04	0.09	0.28	-0.10	0.09	0.32
			M2	-0.55	0.08	0.21	+0.11	0.02	0.07	-0.08	0.05	0.19
			M3	-0.56	0.08	0.20	+0.07	0.02	0.07	-0.08	0.06	0.21
			M4	-0.07	0.01	0.03	+0.15	0.05	0.16	+0.10	0.08	0.30
			M5	-0.37	0.04	0.12	+0.10	0.07	0.22	-0.15	0.06	0.22
			M6	+0.05	0.03	0.08	+0.14	0.03	0.09	+0.01	0.02	0.09
	Cut score *Y*	Raw test score	M1	38.86	1.37	3.63	43.80	1.50	4.73	33.23	1.30	4.83
			M2	33.14	0.51	1.35	37.00	0.93	2.94	27.10	0.46	1.45
			M3	33.86	0.77	2.04	37.10	1.21	3.81	28.00	0.66	2.38
			M4	27.43	1.99	5.26	34.80	2.23	7.05	23.40	1.92	6.08
		Z-scale	M1	+0.63	0.07	0.18	+0.90	0.15	0.47	+0.97	0.22	0.79
			M2	+0.36	0.02	0.06	+0.57	0.04	0.14	+0.29	0.04	0.12
			M3	+0.35	0.02	0.06	+0.54	0.06	0.19	+0.37	0.06	0.21
			M4	+0.03	0.09	0.23	+0.54	0.15	0.49	+0.26	0.07	0.22
			M5	+0.42	0.04	0.11	+0.57	0.07	0.21	+0.37	0.07	0.22
			M6	+0.33	0.04	0.11	+0.49	0.09	0.28	+0.20	0.06	0.19
Resampling (jackknife)				*C**	SE_c*	t_c*	*C**	SE_c*	t_c*	*C**	SE_c*	t_c*
	Cut score *X*	Raw test score	M1	15.56	1.36	11.44	25.31	2.40	10.55	15.57	2.21	7.05
			M2	23.99	1.07	22.42	26.18	0.68	38.50	22.24	0.82	27.12
			M3	22.69	0.99	22.92	25.99	0.59	44.05	22.69	0.60	37.82
			M4	23.31	1.25	18.65	23.50	1.93	12.18	22.23	1.58	14.07
		Z-scale	M1	-0.24	0.06	4.00	+0.09	0.09	1.00	-.13	0.09	1.44
			M2	-0.53	0.08	6.63	+0.14	0.02	7.00	-.15	0.06	2.50
			M3	-0.54	0.08	6.75	+0.05	0.02	2.50	-.06	0.06	1.00
			M4	-0.06	0.02	3.00	+0.19	0.04	4.75	.12	0.08	1.50
			M5	-0.37	0.04	9.25	+0.14	0.07	2.00	-.20	0.06	3.33
			M6	+0.05	0.03	1.67	+0.14	0.03	4.67	-.03	0.03	1.00
	Cut score *Y*	Raw test score	M1	38.88	1.37	28.38	44.00	1.49	29.53	33.21	1.34	24.78
			M2	33.11	0.50	66.22	37.80	0.71	53.24	27.11	0.46	58.93
			M3	33.87	0.77	43.99	37.42	1.22	30.67	28.02	0.66	42.45
			M4	27.45	1.99	13.79	34.22	2.28	15.01	23.39	1.95	11.99
		Z-scale	M1	+0.66	0.07	9.43	+0.90	0.15	6.00	+0.96	0.22	4.36
			M2	+0.35	0.03	11.67	+0.58	0.03	19.33	+0.30	0.04	7.50
			M3	+0.33	0.03	11.00	+0.60	0.06	10.00	+0.44	0.06	7.33
			M4	+0.01	0.09	0.11	+0.56	0.15	3.73	+0.30	0.07	4.29
			M5	+0.45	0.04	11.25	+0.56	0.06	9.33	+0.34	0.07	4.86
			M6	+0.31	0.04	7.75	+0.53	0.08	6.63	+0.25	0.06	4.17

C – cut score (mean);
SE_c – standard error of the cut score C
SD_c – standard deviation of the cut score C
M1 – Basket procedure
M2 – Cumulative Compound method
M3 – Cumulative Cluster method
M4 – ROC-curve method
M5 – Item Mastery method
M6 – Level Characteristic Curves method
t_c – replicability criterion ($t_C > t_{cr\,\alpha/2,\,n-1}$)

Appendix 7

Degree of consistency between the classification decisions

Index	Test	Scale	Method	% of correspondence									
				Raw test score				Z-scale					
				M1	M2	M3	M4	M1	M2	M3	M4	M5	M6
Coefficient of correspondence – k	Test 1 (n_1 = 900)	Raw test score	M1	100	70	73	72	90	84	83	59	88	71
			M2	0.56	100	**97**	69	74	77	77	85	80	98
			M3	**0.60**	**0.94**	100	72	78	80	80	85	84	96
			M4	0.55	0.47	0.52	100	63	87	88	79	81	70
		Z-scale	M1	0.85	0.62	0.66	0.42	100	75	74	63	82	75
			M2	0.73	0.63	0.68	0.76	**0.60**	100	**99**	66	**93**	78
			M3	0.71	0.64	0.68	0.76	0.58	**0.98**	100	67	**92**	79
			M4	0.39	0.72	0.74	0.61	0.45	0.45	0.45	100	70	**86**
			M5	0.81	0.69	0.74	0.68	**0.72**	**0.88**	**0.86**	0.50	100	82
			M6	0.57	0.96	0.93	0.48	**0.63**	**0.65**	**0.67**	**0.74**	**0.71**	100
	Test 2 (n_2 = 277)	Raw test score	M1	100	74	74	64	93	74	75	70	73	68
			M2	**0.63**	100	**100**	**86**	71	92	94	90	92	91
			M3	**0.63**	**1.00**	100	**86**	71	92	94	90	92	91
			M4	0.48	**0.78**	**0.78**	100	58	78	82	77	78	79
		Z-scale	M1	0.88	0.58	0.58	0.40	100	73	74	69	73	67
			M2	0.61	0.88	0.88	0.67	**0.60**	100	**97**	**96**	**100**	**95**
			M3	0.62	0.91	0.91	0.71	**0.62**	**0.95**	100	**95**	**97**	**93**
			M4	0.57	0.85	0.85	0.64	0.56	**0.95**	**0.93**	100	**96**	**97**
			M5	0.61	0.88	0.88	0.67	**0.60**	**1.00**	**0.95**	**0.95**	100	**95**
			M6	0.54	0.86	0.86	0.68	0.54	**0.92**	**0.90**	**0.96**	**0.92**	100
	Test 3 (n_3 = 15370)	Raw test score	M1	100	47	52	30	85	57	67	41	70	43
			M2	0.28	100	**95**	83	40	83	81	91	76	82
			M3	0.32	**0.92**	100	78	45	79	83	86	79	77
			M4	0.16	**0.70**	**0.63**	100	24	73	65	86	60	87
		Z-scale	M1	0.62	0.22	0.25	0.14	100	50	58	34	58	36
			M2	0.35	0.71	0.66	0.53	0.28	100	**91**	**86**	**95**	**86**
			M3	0.44	0.70	0.73	0.43	0.35	**0.86**	100	75	**95**	**78**
			M4	0.23	0.84	0.78	0.74	0.18	**0.73**	**0.61**	100	70	**87**
			M5	0.49	0.63	0.67	0.36	0.30	**0.77**	**0.91**	0.53	100	72
			M6	0.24	0.69	0.62	0.74	0.20	**0.76**	**0.63**	**0.77**	0.55	100

Appendix 8

Criteria for quality: Pair-wise comparison of the analysed methods for setting cut scores

Criterion I – *Range of application*

	M1	M2	M3	M4	M5	M6	Σ
M1	0.5	0.5	0.5	0.5	1	1	**4**
M2	0.5	0.5	0.5	0.5	1	1	**4**
M3	0.5	0.5	0.5	0.5	1	1	**4**
M4	0.5	0.5	0.5	0.5	1	1	**4**
M5	0	0	0	0	0.5	0.5	**1**
M6	0	0	0	0	0.5	0.5	**1**
Σ	2	2	2	2	5	5	**18**

Criterion II – *Statistical complexity*

	M1	M2	M3	M4	M5	M6	Σ
M1	0.5	1	1	1	1	1	**5.5**
M2	0	0.5	1	1	1	1	**4.5**
M3	0	0	0.5	0.5	0.5	1	**2.5**
M4	0	0	0.5	0.5	0.5	1	**2.5**
M5	0	0	0.5	0.5	0.5	1	**2.5**
M6	0	0	0	0	0	0.5	**0.5**
Σ	**0.5**	**1.5**	**3.5**	**3.5**	**3.5**	**5.5**	**18**

Criterion III – *Consistency with the empirical data*

	M1	M2	M3	M4	M5	M6	Σ
M1	0.5	0	0	0	0	0	**0.5**
M2	1	0.5	0.5	0.5	0.5	1	**4**
M3	1	0.5	0.5	0.5	0.5	1	**4**
M4	1	0.5	0.5	0.5	0.5	1	**4**
M5	1	0.5	0.5	0.5	0.5	1	**4**
M6	1	0	0	0	0	0.5	**1.5**
Σ	**5.5**	**2**	**2**	**2**	**2**	**4.5**	**18**

Criterion IV – *Misplacement of the cut scores*

	M1	M2	M3	M4	M5	M6	Σ
M1	0.5	0	0	0.5	0	0	**1**
M2	1	0.5	1	1	0	0	**3.5**
M3	1	0	0.5	1	0	0	**2.5**
M4	0.5	0	0	0.5	0	0	**1**
M5	1	1	1	1	0.5	0	**4.5**
M6	1	1	1	1	1	0.5	**5.5**
Σ	**5**	**2.5**	**3.5**	**5**	**1.5**	**0.5**	**18**

Criterion V – *Standard error of the cut scores*

	M1	M2	M3	M4	M5	M6	Σ
M1	0.5	0	0	0	0	0	**0.5**
M2	1	0.5	1	1	1	1	**5.5**
M3	1	0	0.5	1	1	1	**4.5**
M4	1	0	0	0.5	0	0	**1.5**
M5	1	0	0	1	0.5	0	**2.5**
M6	1	0	0	1	1	0.5	**3.5**
Σ	**5.5**	**0.5**	**1.5**	**4.5**	**3.5**	**2.5**	**18**

Criterion VI – *Significance of the differences between two sequential cut scores*

	M1	M2	M3	M4	M5	M6	Σ
M1	0.5	1	1	1	0.5	1	**5**
M2	0	0.5	0.5	1	0	1	**3**
M3	0	0.5	0.5	1	0	1	**3**
M4	0	0	0	0.5	0	0.5	**1**
M5	0.5	1	1	1	0.5	1	**5**
M6	0	0	0	0.5	0	0.5	**1**
Σ	**1**	**3**	**3**	**5**	**1**	**5**	**18**

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

Felianka Kaftandjieva

Dr. Felianka Kaftandjieva worked at the University of Sofia as an Associate Professor in Educational Measurement and Evaluation from 1997 until her untimely death in 2009. She earlier had a fellowship at Cito and worked at the University of Jyväskylä. She was a consultant in several countries. She was one of the founding members of EALTA and served as Chair of its Membership Committee. Her research interests included Item Response Theory, Standard Setting and Research Design.

This book is a translation of Dr. Kaftandjieva´s second doctorate approved in 2008. She provides an overview of the development of standard setting and reports on her comparative analysis of six standard setting methods that she developed on her own or with some colleagues over a number of years. EALTA is grateful to Dr. Kaftandjieva for her many valuable services to the Association and is happy to have obtained the right to publish this study in a posthumous translation.
EALTA believes that the book will make a substantial contribution to the art and science of standard setting.

Photo: ronsteemers.com

European Association for Language Testing and Assessment (EALTA)