METHODS FOR SETTING CUT SCORES IN CRITERION – REFERENCED ACHIEVEMENT TESTS

COMPARATIVE ANALYSIS OF THE QUALITY OF THE SEPARATE METHODS

Felly Stojanova

Abstract. The different methods lead to different cut scores. 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. The methods for setting cut scores are 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 methods, in terms of their validity (internal and external). One of the hypotheses is that the method with the smallest standard error of the cut score is the Compound Cumulative method.

Keywords: cut scores, criterion-referenced achievement test, basket procedure

Methods

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 indispensable 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, 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.

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

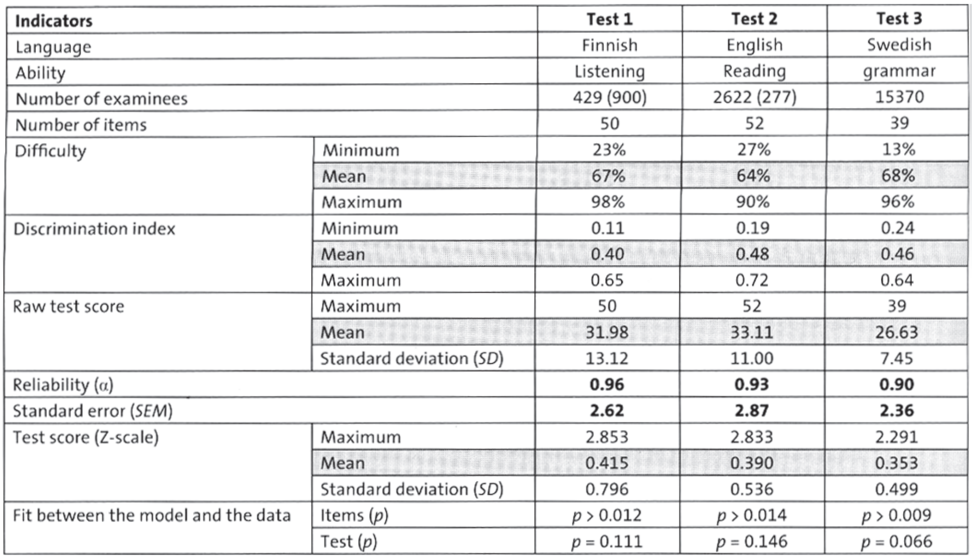
The cut scores, regardless of the method applied, depend to a large extent on the quality of the judgments - whether and to what extent there is agreement between the different judges on the one hand and the judgments and the empirical data on the other. In order to explore the influence of the quality 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.

Instruments

In analyzing 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 (Tl) 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.



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 psychometric characteristics of the test are presented in Table 1 (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 Al 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.

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 1 (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 1, 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 (a = 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.

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 (Al, A2, Bl, B2, CI 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.

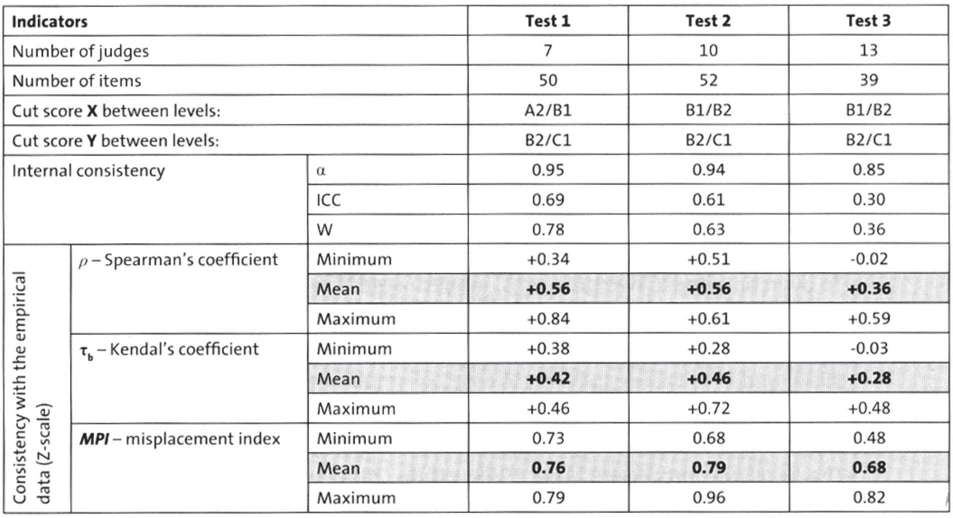
The preliminary instruction of the judges in all three cases was conducted in half a working day and included introduction to the CEFR and the corresponding assessment scale as well as judgment of sample test items in order to increase consistency with the empirical data among the judges.

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.



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

Table 2 presents the values for three of the most frequently used indicators of internal consistency – Cronbach's alpha (a), the intraclass correlation (ICC) and Kendal's coefficient of concordance (Kendall's W). In contrast to the other two indicators, Cronbach's u 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 2 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.

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. 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 confirms that they measure the same characteristic. In this, as could be expected, the relationship between the misplacement index (MPI) and Kendal's coefficient is relatively stronger than between the MPI and Spearman's coefficient.

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).

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 without replacement. 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).

REFERENCES

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

Bontempo, Br. et al. (1998). *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.

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

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

Kier, Fr. (1997). *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.

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

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

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

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

Thompson, Br. (1994). The Pivotal Role of Replication in Psychological Research: Empirically Evaluating the Replicability of Sample Results. *Journal of Personality*, *62(2)*, 157–176.

Verhelst, N. et al. (1995). *One-Parameter Logistic Model: OPLM*. Arnhem, Cito.

White, A. (2000). *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.

Yu, Ch. (2003). Resampling methods: Concepts, Applications, and Justification. *Practical Assessment, Research & Evaluation*, *8 (19)*.

МЕТОДИ ЗА ОПРЕДЕЛЯНЕ НА ПРАГОВИ СТОЙНОСТИ ПРИ КРИТЕРИАЛНИТЕ ТЕСТОВЕ ЗА ПОСТИЖЕНИЯ. СРАВНИТЕЛЕН АНАЛИЗ НА КАЧЕСТВАТА НА ОТДЕЛНИТЕ МЕТОДИ

Резюме. Различните методи водят до различни резултати по отношение на праговите стойности. Дори и за едни и същи експерти праговите стойности са различни в зависимост от различните методи. За някои от методите праговите стойности се различават в голяма степен. Методите за определяне на праговите стойности са методът на сортирането; интегративният комулативен метод; комулативният клъстерен метод; методът на диагностичната крива; методът на минималната загуба и методът на характеристичните криви. Целта на настоящото проучване са психометрични характеристики на методите по отношение на тяхната валидност (вътрешна и външна). Една от хипотезите е, че методът с най-малка стандартна грешка на крайната оценка е интегративният комулативен метод.

Assoc. Prof. Felly Stojanova, PhD