On the Effect of Using Different Scoring Methods for Two Versions of a Test

Martina Hulešová

Abstract: This article presents a study of the effect of a different scoring method on the construct of the Czech Maturita English examination. In particular it focuses on decision consistency made on the basis of the test results and the implications for test fairness and validity of the interpretations of test results. Questions are discussed concerning construct validity, decision consistency and fairness by comparing the test results of two versions of the same test, but with different scoring. The findings show that rescoring causes changes the weights of skills measured by the tests, and thus changes in construct; decision consistency of the tests with different scoring was low, and therefore the interpretation of the results of the two test versions cannot be the same. It was found in this particular case that the students tested do not change their strategies, as they believe that the tests are equivalent and fair, and they are not conscious of the possible consequences of rescoring. On the basis of the results, this article tentatively concludes that introducing different scoring may increase unreliability and cause unfair decisions and judgements of students' ability.

Key words: fairness, construct, scoring method, equivalence

Abstrakt: Změna skórování testových verzí souvisí se změnou váhy ověřovaných dovedností a posunem v definici konstruktu, čímž znesnadňuje interpretaci výsledků testovaných ve dvou verzích téhož testu stejným způsobem. Závěry studie naznačují, že při změně konstruktu je problematické považovat testové verze za paralelní či ekvivalentní a že může docházet k ohrožení validity závěrů a spravedlivosti rozhodování o úrovni dovedností testovaných.

Introduction

In the last 25 years many new high-stakes national examinations have been developed in Europe, and many of them are forced to compromise on best practice due to a variety of internal and external institutional constraints: the availability of resources, political pressure, the non-existence of a national strategy for education, poorly designed policies, etc. (Pižorn & Nagy, 2009; personal communication). In situations where the very existence of the exams or institutions is threatened after every national election, it is very difficult to find stability and resources to conduct research on issues such as fairness or validity, or to apply principles of good practice. Instead of learning from experience, many institutions repeat the same mistakes and poor decisions.

The Czech Maturita exam (the upper-secondary school leaving examination) is a high-stakes exam and its results (especially a pass or fail) influence the lives of the test-takers. Thus, issues such as fairness, test versions' equivalence over a period of years, construct validity, etc. become highly relevant. Changes in the exam format might threaten construct validity and the interpretation of the results. One of these changes was the decision to change the scoring method in 2012.

Ethics, fairness, construct validity

Ethics and fairness are two of the most important issues in language testing. Test providers, developers, and users have to be sure that a test is fair to the candidates. The concept of fairness is closely related to the validity of the interpretation of the test results and to the rationale of the test specifications and test construction process. Striving for fairness and validity of the test results increases in importance in high--stakes contexts.

Validity studies are extensive in the testing literature; however, there are very few practically oriented studies (case studies or validation studies) that discuss the whole process of test validation from the very initial steps of the test development, including the preliminary decisions involving test purpose, design and score uses. Few studies discuss the rationale for selecting a particular scoring method and item weighting, especially for receptive skill testing – organizations usually provide information about how they score their tests, but they do not explain why a particular scoring method is used. The same applies to the literature about item weighting.

Two important concepts of test fairness were introduced by Messick (1995): the 'freedom from bias in scoring and interpretation, and the appropriateness of the test-based constructs or rules underlying decision making' (p. 742). He also suggests that construct validity is a comprehensive concept '... based on an integration of any evidence that bears on the interpretation of meaning of test scores – including content- and criterion-related evidence' (p. 742), with construct representation as its fundamental feature. For him, construct representation refers to the processes, strategies and knowledge involved in the process of task solving, resulting in scores. Test score, for him, are 'an extensible set of indicators of the construct.'

According to Messick (1995), there are two major threats to validity: construct under-representation and construct-irrelevant variance. We consider that varying the scoring or item weighting across test versions is one way in which construct--irrelevant variance is introduced and thus, the construct validity can be affected; we state that if two versions of the same tests are not scored equally, the constructs of these two tests might not be identical and it might not be possible to interpret the constructs in the same way.

Scoring methods

Scoring methods operationalize the meaning of the score and the construct represented by the items. The weight of an item must reflect the construct, or at least it must not add construct-irrelevant elements.

Chapelle (2012) states that “(t)est developers need to provide backing for whatever assumptions underlie the scoring procedures” (p. 26). Alderson et al. (1995) discuss the practical aspects of scoring, such as score form, correction for guessing, and weighting of items or tasks. They define weighting as giving more or an extra value to some items or groups of items (Alderson et al., p. 149), because testers believe those tasks or items are more important for or more representative of the content domain, more difficult or time consuming, or require higher proficiency, and they identify the reasons for weighting test components: components are significant indicators of language proficiency, assess the curriculum content or are particularly time-consuming. However, the authors agree with Ebel and Frisbie (1991) that weighted scoring can be more effectively replaced by adding more items, or using other scoring models, for example, the partial-credit model. According to Jenkinson (1991), scoring and weighting might express the values of the test developer rather than the values of the test stakeholders (p. 1413); if weights are summed to form a total raw score, this raw score can be reached by many different ways. This is important for the construct validity, i.e. for the interpretation of score meaning, as this issue might be a threat to construct validity.

Rotou, Headrick and Elmore (2002) emphasize the importance of a careful selection of the scoring method used for deriving final test scores, since such methods might have substantial effects on score interpretation and subsequently on decisions about test-takers; scores must be interpreted in terms of the construct and the score interpretations must be consistent.

In sum, two scoring methods used for different versions of the same test which yield different results and different (inconsistent) final decisions, are unacceptable in many contexts, but dangerous in high-stakes contexts because of the risk of flawed decisions that affect test-takers.

Research context and research questions

In 2011, a new examination system was introduced. English was one of the compulsory exams. Test specifications valid from 2009 weight each test item equally (1 point). The total weight of a subskill is represented by the total number of items focusing on this subskill, not by assigning more points to a particular item. The internal proportions of subskills were carefully weighted and related to the construct and content described in the test specifications based on the CEFR (2001) descriptors for the B1 level. This scoring method was used in 2010 and 2011 (about 8 test versions). But in 2012, scoring was changed and about half of the tasks were double-weighted.

The Maturita is a high-stakes compulsory exam; thus, issues like parallel form reliability of test versions or test version equivalence, construct validity, and test fairness become highly relevant.

On the basis of the context outlined above, the following research questions were formulated:

Question 1: To what extent do different scoring methods affect the interpretation of the construct being tested – construct validity?

Question 2: To what extent do different scoring methods affect the post-test decision making process in criterion-referenced high-stakes exam[1]?

We hypothesise that:

- The construct changes when item weights change.
- Decision consistency for pass-fail results is low for two test versions when scored differently.
- The results of two versions of the same test cannot be interpreted in the same way when each version is scored differently.
- Students use different test-taking strategies when different scoring is applied.
- The use of different scoring methods for two equivalent test versions threatens the construct validity and the test fairness, as it affects negatively the interpretation of test results and thus the reliability of decisions based on the test scores.[2]

Research design

Participants (a convenient sample of future Maturita test-takers) sat two reading test versions, which were scored by different methods – Method 1 where all items scored one point, and Method 2 where half of the items scored one point and half of the items scored two points. Finally, participants filled in a questionnaire about their test-taking strategies or participated in interviews and observations. Participants' teachers took part as administrators and also filled in a questionnaire about the test-taking strategies they had taught.

Teachers evaluated their students as B1 students. Unfortunately, not all students took both tests and for this reason, the total number of participants decreased to 141 students (right-hand column of Table 1).

Test A is the complete reading subtest from the 2012 Sample Exam. Test B was compiled from two versions of the live tests used in May 2011, to minimize the potential learning effect.

[1] For practical reasons, the research was restricted to the B1 English reading subtest.

[2] The research hypothesis states that ANOVA (analysis of variance) two-way repeated measures will find statistically significant differences given the scoring method, which will also affect decision consistency. Using different scoring methods for two versions (A and B) of the same test leads to inconsistent classification of the test-takers as passing or failing, and simultaneously, the construct of the tests changed in that the interpretations of the results are different when using scoring method 1 and scoring method 2.

Tab. 1: *Structure of the group of participants*

Type of school	Teacher ID	Class ID	Year 4	Year 3	*N* of students	*N* of students (both tests)
Vocational school	MD	1	X		12	12
	ZS	2	X		9	9
	ZS	3	X		8	8
	VK	4	X		14	14
Grammar school	EP	5	X		18	15
Vocational school	AL	6	X		6	6
Vocational school	LK	7	X		24	19
	LK	8		X	11	11
Vocational school	AR	9	X		19	16
	AR	10		X	17	13
Vocational school	EL	11		X	13	10
	EL	12	X		9	8
Total		12	119	41	160	141

Table 2 shows the comparison of the old scoring valid until 2011 and the new scorings applied in 2012. The scoring has remained dichotomous, with 15 out of 25 items double-weighted.

Tab. 2: *Old and new scoring of Tests A and B*

	Task 1	Task 2	Task 3	Task 4	Maximum score
	5 short texts 5 MCQs (4 options)	1 text 10 true/false items	1 text 5 MCQs (4 options)	5 matching items (7 options)	
Old scoring (Tests A1_B1)	5 points	10 points	5 points	5 points	25
New scoring (Tests A2_B2)	5 points	20 points	5 points	10 points	40

Two questionnaires were used (teachers' and students' questionnaire). The former contained closed questions about which tests from the Maturita website teachers had practiced with their students and which test-taking strategies they had taught their students. The students' questionnaire contained closed and open questions asking students which tests they had practiced and what strategies they had used when taking the tests. The main aim was to investigate what kind of test-taking strategies were used during test-taking and whether students use different test-taking strategies in two tests with different scoring methods.

Observations were conducted in four classes with the aim to see what students were doing with the tests while taking them. The observation schedule was very simple, mainly note-taking and describing the observable behaviour, such as reading the information on the title page, underlining in the test booklet, movements signalling thinking, self-correction, moving pages and deciding where to start, etc.

Interviews were conducted with students pre-selected during the observation. The criteria for the selection were their willingness to participate, the quickness or slowness in solving the tests, and some aspects of their behaviour such as constant movement, browsing through the test, etc., which might have indicated the use of test--taking strategies. Interviews were semi-prepared, meaning that if necessary, additional questions were added in order to further illuminate any issue. Interviews were conducted with individual students and in one case due to time constraints, with a small group of respondents. In total, 10 students participated (see Table 3).

Tab. 3: *Techniques used in the study*

Type of school	Teacher ID	*N* both tests	*N* both tests and signed questionnaire	*N* observations	*N* Interviews
Voc. school	MD	12	12	12	3
	ZS	9	1	9	
	ZS	8	8	8	
	VK	14	12	14	4
Grammar school	EP	15	17		
Voc. school	AL	6	6	6	3
Voc. school	LK	19	1		
	LK	11			
Voc. school	AR	16			
	AR	13			
Voc. school	EL	10	12		
	EL	8	8		
Total		141	77	49	10

In order to account for test-order effect, the subjects were divided into four groups and the balanced testing design described in Table 4 was used. Each group received one of the four possible combinations of Tests A and B and scoring Methods 1 and 2.

Tab. 4: *Balanced testing design*

Group 1	Group 2	Group 3	Group 4
A1: Test A Scoring 1 B2: Test B Scoring 2 *N* = 41	B1: Test B Scoring 1 A2: Test A Scoring 2 *N* = 32	A2: Test A Scoring 2 B1: Test B Scoring 1 *N* = 32	B2: Test B Scoring 2 A1: Test A Scoring 1 *N* = 36

Preliminary assumption

A series of assumptions about the tests had to be accepted as valid because gathering evidence of all aspects of validity is beyond the scope of this study.

1. The construct of Tests A and B is based on the definition of reading at the B1 level of the CEFR (2001); the items included in the tests are content-relevant and a representative selection of the construct.
2. The tests are criterion-referenced; their purpose is to measure how well students achieve the aims defined in the Czech curriculum.
3. The cut score at 44% expresses the standard set by the Czech Ministry of Education, whereas a cut score of 65% expresses the mastery level for the B1 standard[3].

Preliminary analyses

Preliminary analyses were conducted with the aim to investigate test order effect, test equivalence, item quality, score distribution and its suitability for parametric analyses.

ALTE Multilingual Glossary of Language Testing Terms (1988, as quoted in Khalifa & Weir, 2009, p. 193) defines equivalent tests as tests that:

> "are based on the same specifications and measure the same competence. To meet the strict requirements of equivalence under classical test theory, different forms of a test must have the same mean difficulty, variance, and co-variance, when administered to the same persons."

This is a very strict definition and very difficult to attain in practice (Taylor, 2004, as quoted in Weir, 2009, p. 193), and given the fact that CERMAT probably does not use an IRT based item bank, it is almost impossible to attain this. The difficulty in attaining the statistical equivalence in practice is also stated in the Manual for Language Test Development and Examining (MLTDE, 2011, p. 82).

To investigate test equivalence, content analysis and a series of statistical analyses were conducted and the results are discussed and considered in relation to the difficulty in attaining test equivalence.

Preliminary analyses – conclusions

Tests A and B resulted in almost identical content, showed sufficient quality, with one exception (Item 12, Test A), which had near-normal score distributions and almost equal variance. There was a statistically significant difference in mean difficulty, but

[3] This cut score was set from a 2011 standard setting project using the Cito variation of the Bookmark method (Verhelst & Hulešová, 2011).

with a small effect size. The sample size was large enough (Ntotal = 141) and the data were independent. Two apparent "outliers" were real data and could not be removed.

Research analyses

ANOVA (analysis of variance) was used to investigate whether different scoring methods really affect the results of the students. Content analysis compared the weights of subskills before and after rescoring. Decision consistency was investigated as an indicator of reliability in CR tests (for 44% and 65% cut score levels). Analyses of teachers' and students' questionnaires on test-taking strategies, how they influence both the way students complete the test and the results, were conducted.

Findings

Questionnaires, observations and interviews: summary

When students received the tests, 20% of them declared they did a quick survey of the entire test; the same number was looking for the information about whether they would be penalized for omitted or incorrect answers. 40% planned the time they needed for parts of the tests; 20% decided to start with other than the first part. Only 4% reported starting with the highest valued items and 13% of students reported they had started with the easiest items, but "easiness" had a different meaning for each of them, as there was no pattern of what the easiest part was since they started with different parts. Students guessed very little, or they made an 'educated guess'. The number of missing answers was quite low, probably because the test was not speeded and it was perceived as relatively easy for most of the students. Students also reported they did not change their answers at the last moment (90%) and they did not copy (95%).

Observations (*N* = 49) and interviews (*N* = 10) did not reveal new or surprising information other than that revealed from the questionnaires. Only a few students went through the test before they started answering it, solving the test mainly in a linear way as they were afraid of omitting something; no signs of misunderstandings or problems with the tests were noticed. However, nobody had read the title page of both tests: they did not read any information, or they read the title page of the first administered test only. They explicitly stated that they had believed that the information on the title page was always the same and they did not think about the consequences of the new scoring. This was in spite of the fact that they were told the aim of the survey and they were informed about different scoring, and that the scoring was less important for them than to do the tests as a whole, and they might use this information in case the test was difficult or timed.

Research question 1:

To what extent do different scoring methods affect the interpretation of the construct being tested (construct validity)?

ANOVA results showed that there is a statistically significant effect of the scoring method on the test results expressed as a percentage correct score, although the effect size is only moderate (Pallant, 2007, p. 208)

Figure 1 provides a graphical representation of the ANOVA results. The graph shows which test the test-takers were given in each group and also the relative positions and differences in the means of both groups and both tests. In order to be sure, a t-test for comparing the means of Group X and Group Y in tests A and B with old scoring (one item – one point) was performed and no significant difference was found (Appendix F).

Knowing that no order effect exists, that there is no difference between groups when tests are scored with the same scoring method and that test-taking strategies do not affect the way students took the tests, it could be concluded that the observed effect is caused by the use of different scoring methods.

Tests A and B were treated as equivalent, although slight differences or issues were found: there is a small difference in the content represented by the items; a statistically significant difference between test means was discovered (test A is slightly easier), but with a very small effect size. Two outliers and one item of poor quality was found in test A (item A12).

If all these findings are included in the interpretation of ANOVA results, a tentative conclusion would be that **the rescoring emphasized the test differences and affected the relative difficulty of the tests for the sample under study, making the rescored tests slightly easier**: Test A was found to be slightly easier than test B at the beginning of this study using the same scoring method (Section 4.2.1.4). After rescoring, Group X took A1 and B2 and the rescored test B2 became more similar to the test A1 in terms of percentage of correct answers. Analogously, Group Y took the originally scored test B1 and the rescored test A2 and the difference in the percentage of correct answers became larger.

Research question 2:

To what extent do different scoring methods affect the decision making process in the criterion-referenced high-stakes exam?

Decision consistency analysis focused on the extent to which the two test versions consistently classify students into pass and fail categories (masters or non-masters).

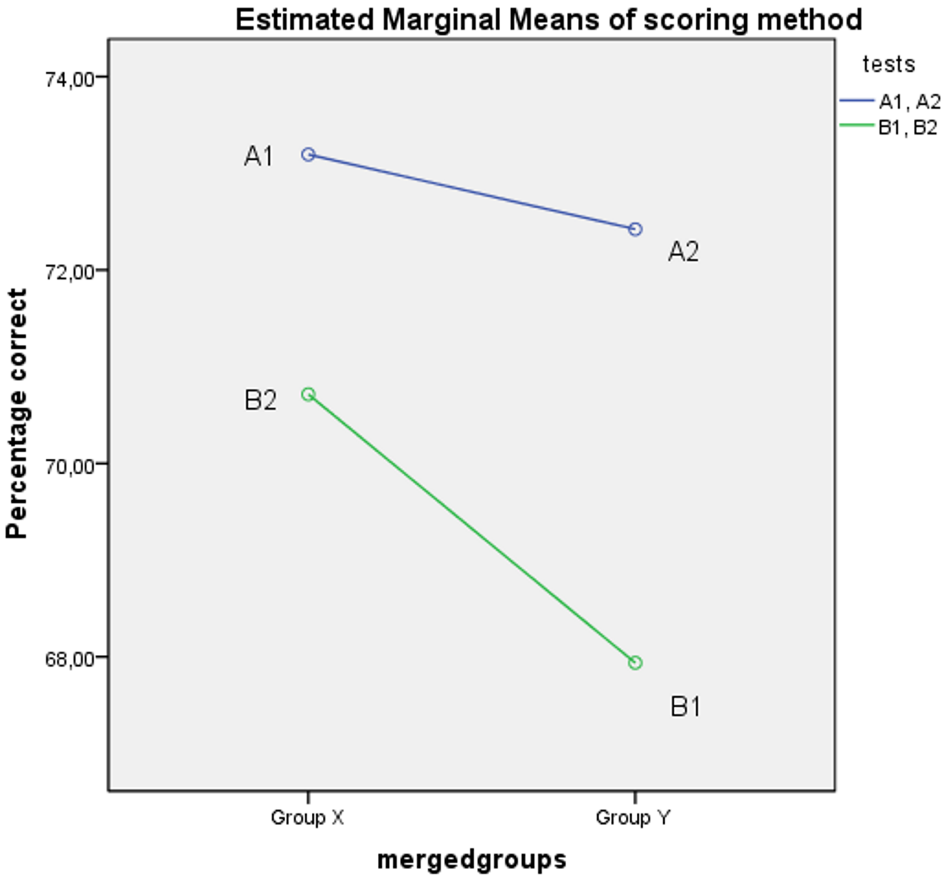

Fig. 1: *ANOVA plot for scoring method effect*

The combination of original and rescored tests was used and Kappa[4] analysis was performed.

The interpretation of the Kappa coefficient is complicated by the fact that other factors can influence its magnitude. Although Landis and Koch (as quoted in Sim & Wright, 2005) proposed the frequently used standards for strength of agreement for the Kappa coefficient (0 = poor, .01–.20 = slight, .21–.40 = fair, .41–.60 = moderate, .61–.80 = substantial, and .81–1 = almost perfect), Sim and Wright (2005) opine that these and similar criteria are arbitrary and the interpretation is incomplete if all the factors that influence the Kappa coefficient are not reported and discussed. For

[4] Kappa – one of the indices of consistency.

this study, three relevant factors were identified: prevalence, bias and probability of occurrence of the categories.

Prevalence expresses how much the proportion of agreements on the master classification differs from that of the non-master classification (Sim & Wright, 2005). A high prevalence index means that chance agreement is also high and Kappa is lower accordingly. A high prevalence index can be seen at the 44% cut score level: in Group X, the prevalence index is .84, in Group Y is .83, whereas the prevalence index in Groups X and Y at the 65% cut score level is considerably lower: .32 and .24 respectively, which means that with the low cut score level almost all students are classified as masters and the category of masters is dominant.

Bias is a kind of a complementary index to the prevalence index and expresses the extent to which the classifications disagree on the proportion of masters or non--masters (Sim & Wright, 2005). At the 44% cut score level, the bias-indices are 0 (Group X) and .01 (Group Y), which is logical as there is a very low number of failed students. At the 65% cut score level, where the proportion of failed students increased, the bias index (its absolute value) is slightly higher, but still very low: .05 in Group X and .17 in Group Y.

Another factor that influences the magnitude of Kappa is the probability of occurrence of the categories. Kappa is usually higher if the occurrence of all categories is equally probable (Sim & Wright, 2005). Here we have only two categories, master and non-master, and given the criterion-referenced nature of the tests and the overall ability of the sample of test-takers, equal probability of occurrence of masters and non-masters cannot be expected – more masters than non-masters are expected for both cut score levels, but especially for the 44% cut score level. This is also confirmed by the results showing that the 44% cut score has almost no power in classifying test-takers into the master and non-master category; the Kappa coefficient related to the 44% cut score is very low; the standard error associated with Kappa and the prevalence index are extremely high.

This study tentatively concludes that scoring method has a small observable effect on decision consistency (master vs. non-master classification) when the 44% cut score is applied, regardless of whether Test A or Test B is used. When the 65% cut score is applied, the number of failed student increases and greater differences can be seen. **It might indicate that there is an interaction between the effect of the cut score, the test version, and the scoring method.** If the Kappa values of Group X and Group Y at the 65% cut score level are compared, it can be seen that while the Kappa for Group X increased substantially, the Kappa for Group Y is substantially lower and it even decreased when compared to the Kappa of Group Y at the 44% cut score level.

This study tentatively concludes that **scoring method influences the decision consistency and also has a small observable effect on the difference in difficulty be-**

tween the two tests: this effect is almost imperceptible at the low cut score level, but it becomes relevant when the new scoring and higher cut score are applied: test A1, when rescored as A2, becomes even easier, and the differences in relative difficulty or percentage correct are accentuated (see Figure 1). If the higher cut score of 65% is used, fewer students pass; therefore, the consequences of rescoring are more visible than if we do not rescore or do not apply the 65% cut score. Found Kappa values, their changes at different cut score levels and all three factors (prevalence, bias and probability of occurrence) confirm this conclusion. The difference between the observed Kappa and 1 (the maximum theoretical value of Kappa), which indicates the total unachieved agreement beyond chance (Sim & Wright, 2005), is rather large in all four Kappa values. Standard errors are very high, especially at the 44% cut score level.

The content analysis provided another view on the test versions' comparability. Only three judges did the analysis, but their agreement was very high. The results of the content analyses support the findings. Rescoring changed the internal weight proportions of skills emphasizing search reading and scanning over global reading (from 60:40 to 75:25). It is also probable that this change caused even small differences in the difficulty of the originally scored tests (as observed by PASW and WINSTEPS[5] analyses and t-tests) to become higher after rescoring.

Conclusion

The results suggest that applying different scoring methods to the same versions of a test might cause a substantial shift in the internal proportion of the weights of the skills that constitute the construct being measured. Consequently, the same test- takers taking two tests might achieve different results expressed as a raw score or percentage correct due to the different scoring method. In the light of these two findings, it can be argued that the performance of test-takers taking two versions of the same test, but with different scoring, cannot be interpreted in the same way due to the effect of the scoring method on the construct to be measured and on the observed performance. The change found in the construct of tests A and B, and thus, in the interpretation of the results, represents a threat to construct validity, which confirms Messick's emphasis on construct representation as one of the basic features of the construct validity evidence (Messick, 1995), and casts doubts on the meaningfulness of weighting items when improvement of reliability and validity of test scores is pursued (Ebel & Frisbie 1991; Alderson et al., 1995).

Thus decision consistency is rather low, and test fairness seems to be threatened. It can be suggested that direct consequences of this inconsistent decision-making process are more clearly visible when the cut score is set around the measures of central tendency (mean and median). If the cut score is too low (e.g. 44%), it loses

[5] PASW and WINSTEPS – statistical software.

its meaning as a functional borderline between master and non-master categories, as it does not distinguish well between those categories. If a test is not consistent or reliable, the interpretation of its results cannot be fair and valid.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Chapelle, C. (2012). Validity argument for language assessment: The framework is simple…, *Language Testing* 29, 19–27.

Council of Europe. (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. New Jersey: Prentice Hall.

Jenkinson, C. (1991). Why are we weighting? Critical examination of the use of item weights in a health status measure. *Social Science & Medicine* 32, 1413–1416.

Khalifa, H., & Weir, C. (2009). *Examining reading*. Cambridge: Cambridge University Press.

Manual for language test development and examining. (2011). Council of Europe. Retrieved December 12, 2012. Available from http://www.coe.int/t/dg4/linguistic/ManualtLangageTest-Alte2011_EN.pdf

Messick, S. (1995). Validity of Psychological Assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist* 50(9), 741–749.

Pallant, J. (2007). *SPSS survival manual*, 3rd edition. McGraw-Hill Education.

Pižorn, K., & Nagy, E. (2009). The politics of examination reform in Central Europe. In Alderson, J. Ch. (Ed.). *The Politics of Language Education: Individuals and Institutions*. Bristol: Multilingual Matters.

Rotou, O., Headrick, T. C., & Elmore, P. B. (2002). A proposed number correct scoring procedure based on classical true-score theory and multidimensional item response theory. *International Journal of Testing,* 2(2),*131–141*.

Sim, J., & Wright, Ch. C. (2005). The Kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy,* 85, 257–268.

Verhelst, N., & Hulešová, M. (2011). *Standard setting in the national examination of English in the Czech Republic*. Retrieved November, 13, 2012, from www.promz.cz/download/1404034454/?at=1

Bionote

Martina Hulešová, e-mail: mhulesova@volny.cz, AJAT – Association of Language Testers in the Czech Republic

Graduated in Spanish philology (Charles University), taught Spanish for 16 years. Graduated from the University of Lancaster (Language Testing). Worked for Maturita project for 11 years. Lecturer, consultant in language testing. PhD student at MUNI (Brno). Researcher at Research and Test Centre (ÚJOP UK).