Situating language tests in relation to the CEFR

Jana Bérešová,
Trnava University, Slovakia
jana.beresova@truni.sk

Abstract

The paper outlines the process of aligning English tests to the CEFR focusing on the stages recommended by the Manual. Beginning with the training of English teachers to interpret the CEFR levels to exemplar test items and tasks, the process of aligning the school-leaving examination tests to the CEFR was based on test items measuring receptive skills (listening and reading) and those measuring the ability to use grammar and vocabulary. Using a multiple linear regression analysis, a high degree of correlation determined the relative importance of the language in use score to the total score.

Carried out in 2014, the research referring to the comparison of teacher´s judgements of test-takers´ performances and test-takers´ testing scores confirmed our assumption related to a gap between teaching and testing. Both the official scores of the test-takers and teachers´ judgements will be presented and commented on. The research has proved that teachers should be trained how to construct a good test and design good items, which is likely to be reflected in their teaching.

Keywords: CEFR, Manual, validation, quality, fairness, teachers´ judgements, test scores

Introduction

The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR), officially published in 2001 with the influential idea of using the descriptor scales to profile the content of courses, assessments and examination, has become very significant for language teaching, having impact on teacher education, syllabus and course design and testing. As Brian North (2004) states the CEFR descriptors offer a practical, accessible tool that can be used to relate course and/or examination content to the CEFR levels, and to train teachers, assessors, and item writers in a standad interpretation of the CEFR levels.

The first impact of the CEFR (2001) is noticeable in many European countries, in which the descriptor scales and illustrative examples were and are still used in curricula design to define expected language proficiency. Despite the fact that different countries generally test different things and each result is reported in terms of the achievement in that particular assessment, most of these countries expressed their wish to link their tests (preferably those designed for the national examinations) to the CEFR. This tendency to relate the local examinations to the CEFR caught the attention of the Language Division of the Council of Europe, which resulted in the publication called Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): A Manual (2009), and local test developers became eager to start the process of linking their tests to the CEFR. A great deal of effort has been done mostly in testing English in local context.

The primary goal of language learning – to foster communicative competence, or the ability to communicate effectively and spontaneously in real-life settings – made the English test designers and item writers embrace a communication perspective of language. As the CEFR promotes an action-oriented approach, most European countries focus on testing receptive and productive skills. Testing both oral/written production and/or interaction is still at stake as many countries cannot ensure objectivity and have problems with reliable and valid marking criteria. Therefore they base their high-stake decisions mostly on the scores achieved in the tests related to testing receptive skills. On the other hand, language learners are generally better at receptive skills rather than productive skills. A lot of work still is to be done in testing productive skills.

Relating tests to the CEFR

The Manual (2009) was written with the aim of helping the providers of any examinations or tests to develop, apply and report transparent and practical procedures in order to link their examinations/tests to the CEFR. The document presents five inter-related sets of procedures which are to be followed to be able to provide both theoretical and practical evidence in order to validate the claim.

These five steps of procedures (Familiarisation, Specification, Standardisation/Benchmarking, Standard-setting and Validation) seem to be in a linear progress, however, the stages related to both the specification process and the standardisation process are highly recommended to start with the familiarisation activities and validation is not an ultimate verdict of the linking process, it is a process of quality monitoring and should be followed from the very beginning to the very end (Noijons et al., 2011).

As the linking process is a team project, those who are involved in this process are expected to have an in-depth knowledge of the CEFR, its descriptors and illustrative samples. These panellists should complete **familiarisation** activities, proposed in the Manual (2009) or the Highlights from the Manual (Noijoins et al., 2011) in order to achieve a high level of familiarity, which will help them to make decisions later.

The **specification** stage involves stating what is and what is not assessed in the examination, and what level of achievements is expected (North, 2004), therefore it can be considered a self-audit of what the examinations cover, involving content and task types, however, it can serve as a report related to previously administered test forms. In addition, this stage has a certain awareness-raising function that can influence the quality of the examination concerned in the future.

The **standardisation** training facilitates the implementation of a common understanding of the common reference levels, during which CEFR illustrative samples for spoken and written production are used, while the **benchmarking** process is based on the judgements referring to performance samples from the test to be benchmarked to the levels that were intended in designing the test. As North (2004) states productive skills are easier to work with as the panellist can see the performances being evaluated and can relate them directly to relevant CEFR descriptors.

Standard-setting is a process of establishing a decision related to allocating the test-takers to one of the CEFR levels, taking into account their performances in the examination. This takes the form of deciding on cut scores or borderline performances. North (2004) emphasizes that in standard-setting, initial estimates of the level of difficulty of an item often bear a limited relationship to the actual difficulty in practice.

In high-stake testing, empirical **validation** of this standard-setting is a requirement. It involves the collection and analysis of data on test scores and two aspects to empirical validation are recommended to follow: internal validations, which is concerned with the quality of the test, and external validation (the provisional conversion of test scores to CEFR levels). North (2004) proposes several ways of data analysis, using simple correlation functions in Microsoft Excel and a couple of Microsoft Word tables. The Manual (2009) recommends that data analysis should be provided by professional statisticians who usually use Item Response Theory (IRT) which focuses on the construct to be measured.

A good linking process requires a quality examination (content validity), a pilot, a pretest and psychometrics.

Teacher´s judgements of item difficulty

In 2008, a group of language professionals (15) trained by Dianne Wall, were asked for rating their students´ performances and their judgements were compared with the official statistical data. The results of this research were analysed and discussed in the article *The Impact of the Common European Framework of Reference on Teaching and Testing in Central and Eastern European*

Context (Bérešová, 2011). The conclusions concerned the gap between the teachers´ judgements and the official scores achieved by the students as the former ones were more consistent while the latter ones reflected the real achievements of the students in English test B2. In both listening and reading sections of the tests the scores proved that the techniques used in receptive skill tasks had been deployed the language in use section revealed that the teachers´ expectations had been higher than the actual students´ results.

In 2014, the same group of panellists was addressed to participate in a similar study, but this time it is necessary to conclude that despite the fact that the teachers´ judgements were consistent, they were higher than the real scores. The teachers admitted that each year the students complained about the task and item difficulty in listening comprehension, but analysing the achievements of students from the previous years, they had been influenced in their judgements by better results from those years and had overestimated their performance.

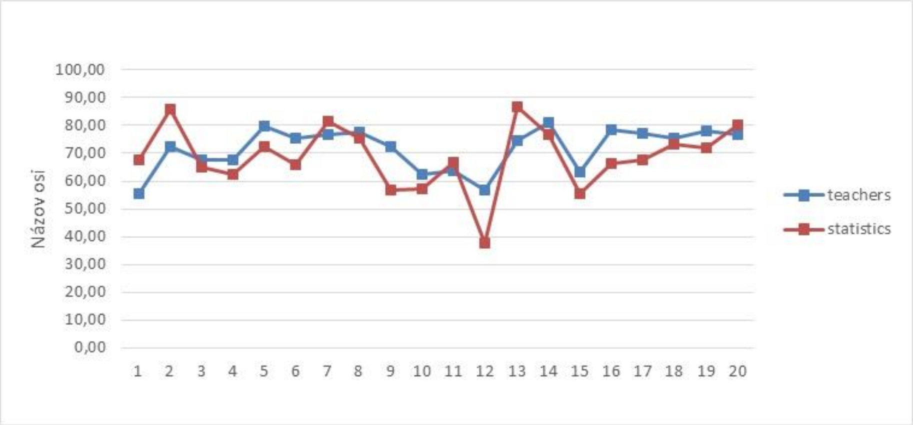

Figure 1 Listening 2014: Teachers´ judgements versus students´ scores

Despite the fact that the CEFR promotes an action-oriented approach, grammar and vocabulary are still tested in many European countries - either in the reading section or in the language in use section. English test B2 in the Slovak leaving-school examination consists of three tasks: 20 multiple-choice items, 10 word formation items and cloze test (10 items). Due to teaching being still based on grammar, teachers´ expectations were sometimes much higher than real performances and students´ problems referred to the selection of correct words from the distractors (items – 23, 25, 29, 34). A good item is one which discriminates strong students from weak ones, demonstrated with problematic items 23 and 34.

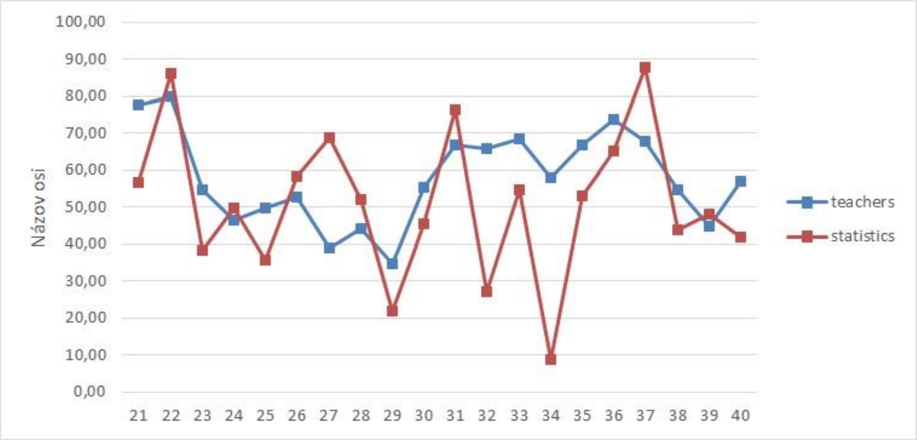

Figure 2 Language in use (multiple-choice): teachers´ judgements versus students´ scores

Analysing the data achieved in the word formation task, it is necessary to comment on the low ability of the students to form new words from the word stems. The results are not satisfactory as the task is rather productive than receptive (multiple-choice), and the learners were required to use the language properly. Many students still have problems to infer proper information from the context and they commented on their failure complaining that there had not been sufficient focus on word formation processes in their English classes. The most difficult item (42) was based on forming the expression *inconvenient* from the base (convenience).

The most surprising gap between teaching and testing appeared in cloze tests, in which the results were poor and teachers´ judgements were vastly overrated. This pragmatic expectancy grammar task forces the students to integrate their knowledge of grammar, meaning and pragmatic use to complete the task. If the students are trained to control their grammar in isolated utterances without context, they are not prepared for controlling the grammatical patterns in this gap-filling task. The most problematic items were 52, referring to the phrasal verb *try out*, item 54 related to the preposition *on* in the expression *on the move* and 57 referring to negative particles *not/never*.

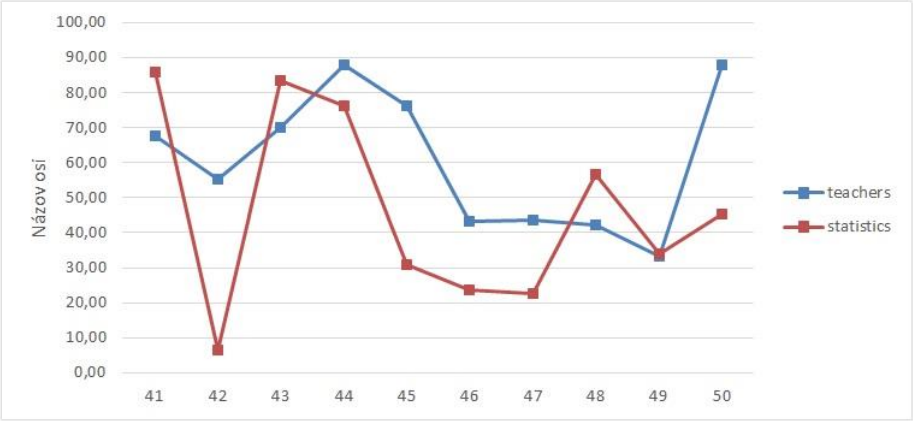

Figure 3 Language in use (word formation): teachers´ judgements versus students´ scores

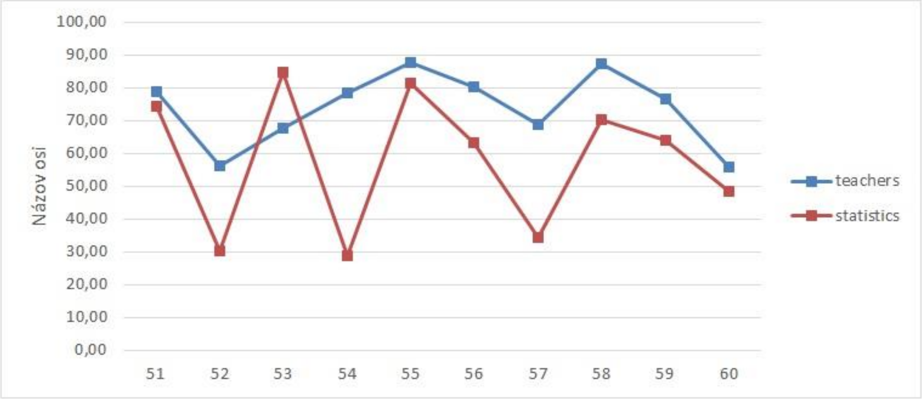

Figure 4 Language in use (cloze test) 2014: teachers´ judgements versus students´ scores

Reading tasks measured comprehension in two closed-item tasks and one open-ended task. The teachers´ expectations were consistently lower in those tasks where the students had been expected to find and comprehend main ideas (matching) and specific information (true/false).

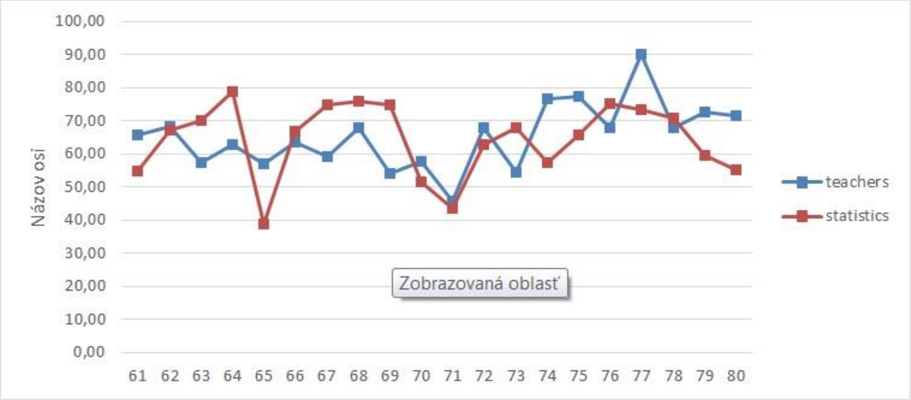

Figure 5 Reading 2015: teachers´ judgements versus students´ scores

Analysing the psychometric data, it can be claimed that English test B2 is a reliable measurement tool with the data as follows: the mean (total) – 64.2, the mean (listening) – 84.2, the mean (language in use) – 48.8, the mean (reading) – 59.6, the standard deviation (Cronbach´s alpha) - .922, although the histograms related to testing receptive skills suggest that the test was not difficult for the tested population.

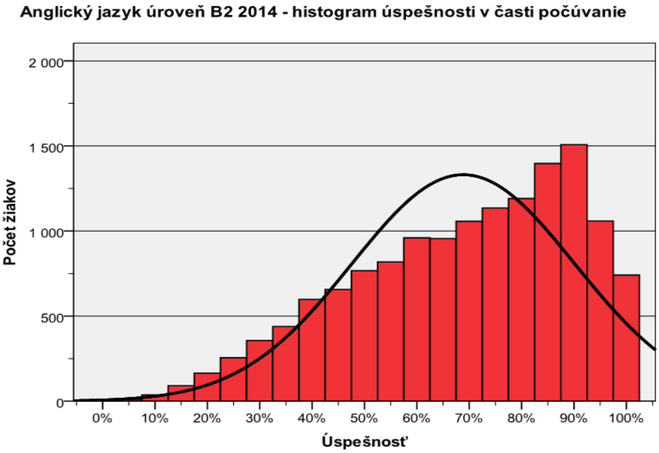

Figure 6a Distribution of students´ receptive skills scores - listening

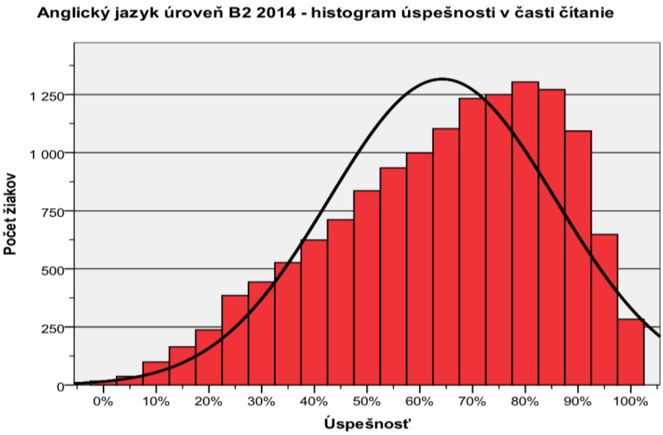

Figure 6b Distribution of students´ receptive skills scores – reading

Analysing the distribution of total scores and comparing its shape with the distribution of scores achieved in the language in use section, they seem to be more similar and the correlation between students´ language in use scores and total scores is strong.

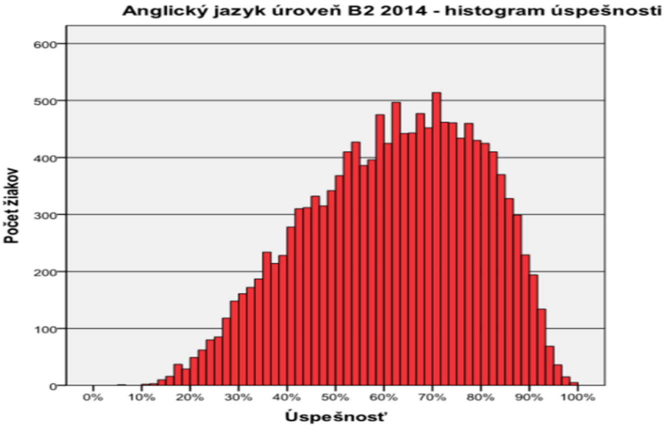

Figure 7a Distribution of students´ total scores and scores achieved in language in use

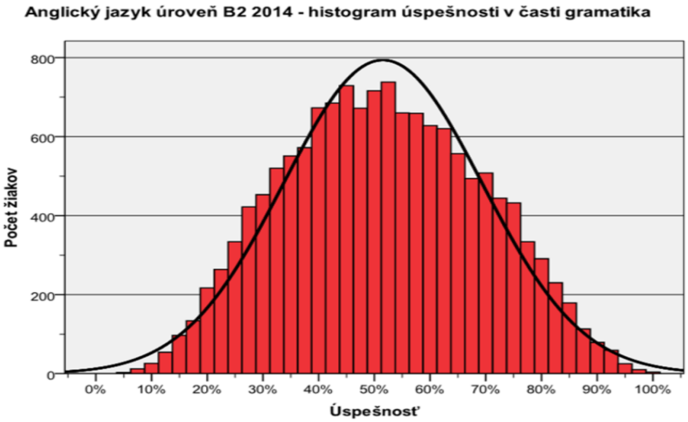

Figure 7b Distribution of students´ total scores and scores achieved in language in use – grammar

The histogram related to language in use scores shows that the distribution of scores can be considered normal (bell-shaped) and this section of the test significantly differentiated successful students from less successful and weak students.

Conclusion

The difference between test development in Slovakia in previous years and last year is significant, as the prepared versions of the tests were pre-tested and item-writers could replace those items which had not worked properly in the pre-testing phase.

The issues that seem to be still critical are the students´ comments on teaching English at secondary schools, which is considered still traditional and many teachers focus on testing skills such as listening and reading in the last year of secondary-school education. Grammar is still a priority of many teachers who test particular grammatical structures in isolated sentences, and their students are not exposed to tests in which grammar is tested in context.

In Slovakia, the discussion related to situating English tests in relation to the CEFR discovered many drawbacks, which should be diminished or removed, such as low cut scores (33%); insufficient objectivity, reliability and validity related to marking the writing papers; insufficient pre-service and in-service teacher training focusing on assessment; subjective traditional marking without using any marking criteria.

A traditional approach to teaching English influences language competence of the students who are not eager to be exposed to English outside English class as reading aloud and translating the English coursebook texts into Slovak demotivates them and English seems to be very difficult for them as they cannot use it for real-life situations.

References

BACHMAN, L. F. (1995). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

BACHMAN, L. F. (2005). *Statistical Analyses for Language Assessment.* Cambridge: Cambridge University Press.

BÉREŠOVÁ, J. (2011). The Impact of the Common European Framework of Reference on Teaching and Testing in Central Eastern European Context. *Synergies Europe*, 6(2011), 177-190p.

Common European Framework of Reference for Languages: Learning, teaching and assessment 2001. Cambridge: Cambridge University Press.

NOIJONS, J., BÉREŠOVÁ, J., BRETON, G., & SZABÓ, G. (2011). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). Highlights from the Manual.* Strasbourg: Council of Europe.

NORTH, B. (2004). Relating assessments, examinations, and course to the CEF. In Morrow, R. (Ed.), *Insights from the Common European Framework* (p. 77-98). Oxford: Oxford University Press.

Relating Language Examinations to the Common European Framework of Reference: Learning, Teaching, Assessment (CEFR). A Manual. (2009). Strasbourg: Language Policy Division.

Resumé

Príspevok sa zaoberá prehľadným vstupom do jednotlivých na seba nadväzujúcich postupov, ktoré odporúča dokument *Relating Language Examinations to the Common European Framework of Reference: Learning, Teaching, Assessment (CEFR). A Manual* (2009), vydaný Radou Európy s cieľom poskytnúť zostavovateľom testov možnosť vyvinúť, aplikovať a zreferovať transparentné a praktické postupy zamerané na nastavenie lokálnych jazykových testov a skúšok na Spoločný európsky referenčný rámec (SERR). Nastavenie testov na SERR si vyžaduje tímovú prácu. Účastníci daného procesu musia prejsť

cez fázy familiarizácie, špecifikácie, štandardizácie (benchmarkingu) a stanovenia noriem tak, aby validácia bola procesom monitorovania kvality, a nie záverečným verdiktom o kvalite procesu.

V príspevku sa zameriavame na porovnanie hodnotenia výkonov maturantov ich učiteľmi v externej časti maturitnej skúšky z anglického jazyka na úrovni B2 s oficiálnymi výsledkami žiakov v školskom roku 2013/2014. Učitelia podhodnotili výkony žiakov, opierajúc sa o svoje skúsenosti z predchádzajúcich rokov, kedy sa žiaci sťažovali na náročnosť časti počúvanie s porozumením. I napriek výučbe zameranej na gramatiku a slovnú zásobu, učitelia vo všetkých troch úlohách časti zameranej na praktické používanie gramatiky a slovnej zásoby (v úlohe s viacnásobnou voľbou odpovede, tvorbe slov a v cloze teste) hodnotili výkony žiakov konzistentne a výrazne nadhodnotili výkony žiakov. Zlyhanie niektorých žiakov stredných škôl je následkom testovania gramatických javov a lexikálnych jednotiek izolovane, nie v kontexte a v ich reálnom používaní. Čítanie a prekladanie na hodinách anglického jazyka demotivuje žiakov, obmedzuje ich skúsenosť s cieľovým jazykom a zbytočne vyvoláva pocit neovládania jazyka pre jeho náročnosť.

Kľúčové slová: CEFR, príručka, validácia, kvalita, objektívnosť, učiteľov úsudok, test, skóre.

Contact

Jana Bérešová
Trnavská univerzita v Trnave
Hornopotočná 23, 918 43 Trnava
jana.beresova@truni.sk