

GAPS BETWEEN LEARNERS' PERFORMANCES AND TEACHERS' EXPECTATIONS

Jana Bérešová

Department of English Language and Literature, Faculty of Education, Trnava University in Trnava,
Priemyselná 4, P.O. BOX 9, 918 43 Trnava, Slovakia

Abstract

The paper outlines a study related to the impact of regular teachers' judgements of students' performances, which resulted in a certain shift from the previous approach of teachers to judging their students' performances (2008) to the current ones (2013, 2014). While in 2008, teachers overestimated learners' performances, taking into account their experience from classroom testing rather than high-stakes testing, in recent years they formed their opinions more carefully. While classroom testing is based on an immediate reflection of material gone over in class, high-stakes testing measures the ability to use English in real life. The paper is based on the official statistical data and compared with the data achieved from the trained teachers.

Key words: statistical analyses, classroom testing, high-stake testing, learners' performances, teachers' judgements

INTRODUCTION

Communicative language teaching focuses on both the functional and structural aspects of language, combining these into a more fully communicative view (Littlewood 1994). Language learners are expected to develop skills that will enable them to process the received information and produce an appropriate response. To be able to use a foreign language efficiently, learners need to acquire both a repertoire of linguistic items as well as strategies for using them in real-life situations. Canale and Swain (1980) influenced language testing by broadening Hymes's (1972) appropriateness of language use (whether something is possible, feasible, appropriate in context and ever actually performed), completing linguistic competence with sociolinguistic competence, discourse competence and strategic competence.

The term communicative language ability provided by Bachman (1995) as ability encompassing both knowledge, or competence, and the capacity for implementing that competence in appropriate, contextualized communicative language use has influenced the practical development and use of language tests: the nature of measurement, the contexts that determine the use of language tests, the nature of communicative language abilities, and the nature of testing methods that are used to measure them.

Many international testing centers provide various tests to measure language competence when language learners need to prove their language acquisition. Therefore current summative testing philosophy is related to a wide variety of assessment batteries that cover both productive and receptive skills, a range of assessment tasks, and increase attention to the communicative properties of tests.

In the late 1990s, Slovakia as many other Central and Eastern European countries being involved in piloting the first draft of the Common European Framework of Reference for Languages: Learning, teaching, assessment (2001) started the reform of the school-leaving examination to change the traditional approach of testing structural aspects of foreign languages. This effort resulted in newly designed tests, officially recognized by the Ministry of Education in 2004. The reform was induced by foreign language teachers who were not satisfied with the imbalance between teaching and testing. From the very beginning, English teachers were intensively trained by language experts from foreign universities and reputable testing centers, mainly from Great Britain.

When the first draft of the Manual: Relating Language Examination to the Common European Framework of Reference for Languages: Learning, teaching assessment (CEFR), published by the Council of Europe in 2008, was available, the country was willing to go through the process of situating their school-leaving language examinations to the CEFR. During the training sessions, the teachers were trained in judging their students performances within standard-setting procedures.

1. TEACHERS' EXPECTATIONS VERSUS OFFICIAL SCORES

Standard-setting is the process of determining a cut-score for a test. The methods for setting standards rely on the subjectivity of experts' judgements, which can be exercised in relation to test content or performance of test takers. Since 2008, trained English teachers have been invited for standard-setting sessions. While being trained, they are asked to judge performances of their students.

In 2008, a group of language professionals (15) trained by Dianne Wall, were asked to rate 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* (Béřešová 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.

1.1 Listening Comprehension

In the late 2000s, listening used to be considered the most difficult skill as there were not enough opportunities to practise listening outside classes in those times, so teachers' expectations were lower than real students' performances.

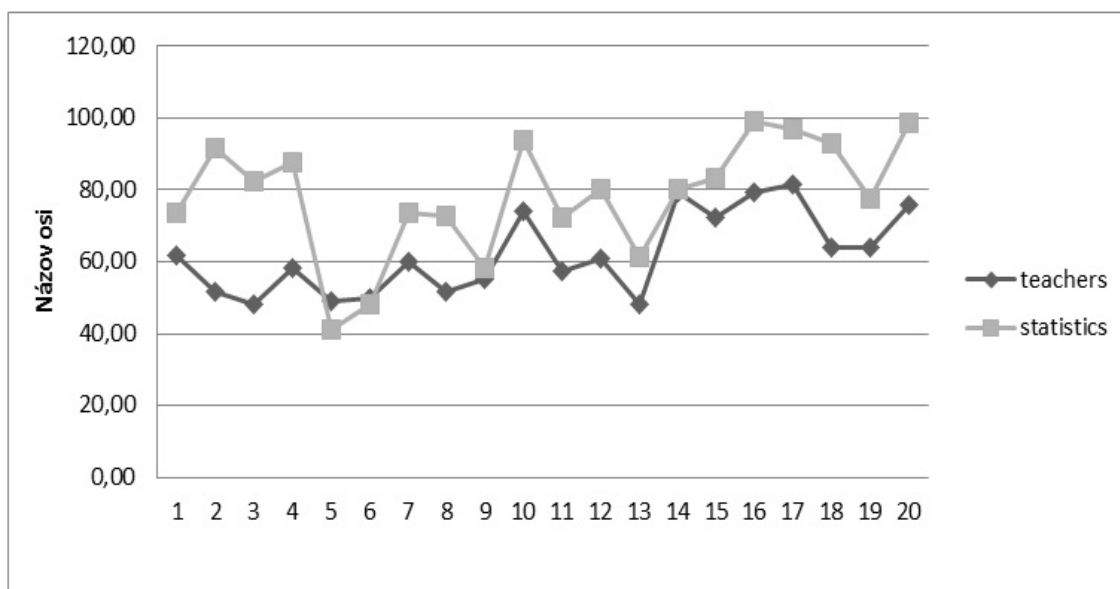


Figure 1 Teachers' judgements versus test scores in Listening 2008

In 2014, the same group of panellists was invited 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. Currently, teachers use various materials to practice listening, and students can listen to recordings at home as each workbook is complemented by a CD for practising listening more intensively. What is more, teachers use a lot of supplementary materials and recommend learners

recordings which are available on the internet and might be useful for training listening to various accents, for example.

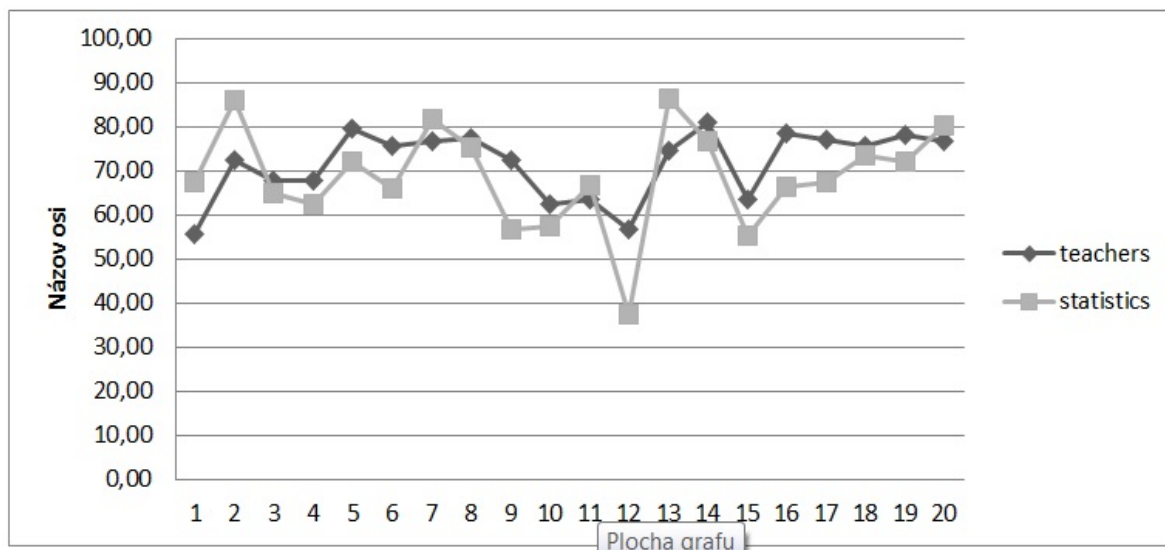


Figure 2 Teachers' judgements versus test scores in Listening 2014

The number of available materials for listening misled the teachers in their judgements. They expected that their students had practised listening at home, using modern technologies. The students admitted that they used modern technologies for their interests rather than for improving their language competence. The second reason influencing teachers' decisions was that Part 3 (14-20) had been replaced from the former one (gap-filling) into a new one (matching). While being marked, the gap-filling task caused a lot of problems related to spelling inconsistency of markers, so to preserve objectivity, the task has been changed. Whereas, listening texts 2008 were predominantly recorded by British native speakers, listening texts 2014 were read by a wide range of native speakers from different English-speaking countries.

1.2 Language in use

Despite the fact that there is still a lot of discussion whether grammar and vocabulary should be part of communicative language tests, which are planned to be linked to the CEFR as the CEFR supports an action-oriented approach, the school-leaving examination is a summative test based on the national curricula and teachers insist on testing grammar and vocabulary as they invest their energy in paying systematic attention to functional as well as structural aspects of language. The language in use section of the test consists of three tasks that are regularly changed. Since a bad impact of testing on teaching might be that teachers force students to spend more time on practising testing techniques rather than learning grammar and vocabulary for real-life purposes. The presented results were achieved in the multiple-choice task as we did not want to compare tasks in which different testing techniques were and still are used. However, multiple choice tasks can be disputable from the validity point of view, there are other drawbacks of using them such as negative backwash effect on teaching or the risk of guessing.

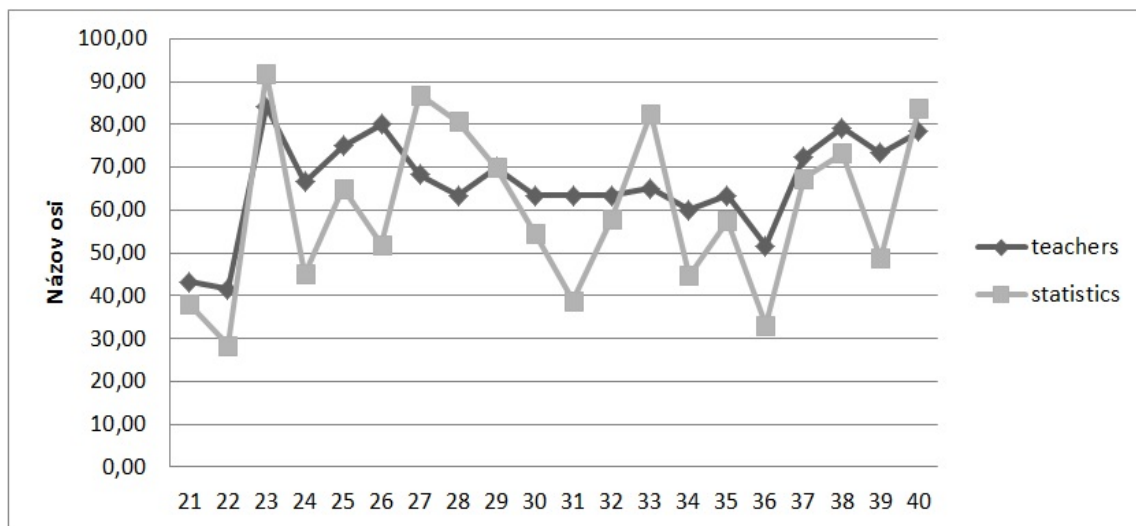


Figure 3 Teachers' judgements versus test scores in Language in Use (multiple-choice) 2008

Both in 2008 and 2014, the trained teachers tried to focus on grammar and vocabulary items specifically, as they rely on their teaching practise and are able to predict which items might be problematic for their students. Regardless of the fact that their estimates are higher than testing scores, some items (11, 3, 25, 26, 29, 40) have graphically similar illustrations. In items 20 and 32 their judgments were very close to the statistical measurement. There is one exact coincidence (item 9).

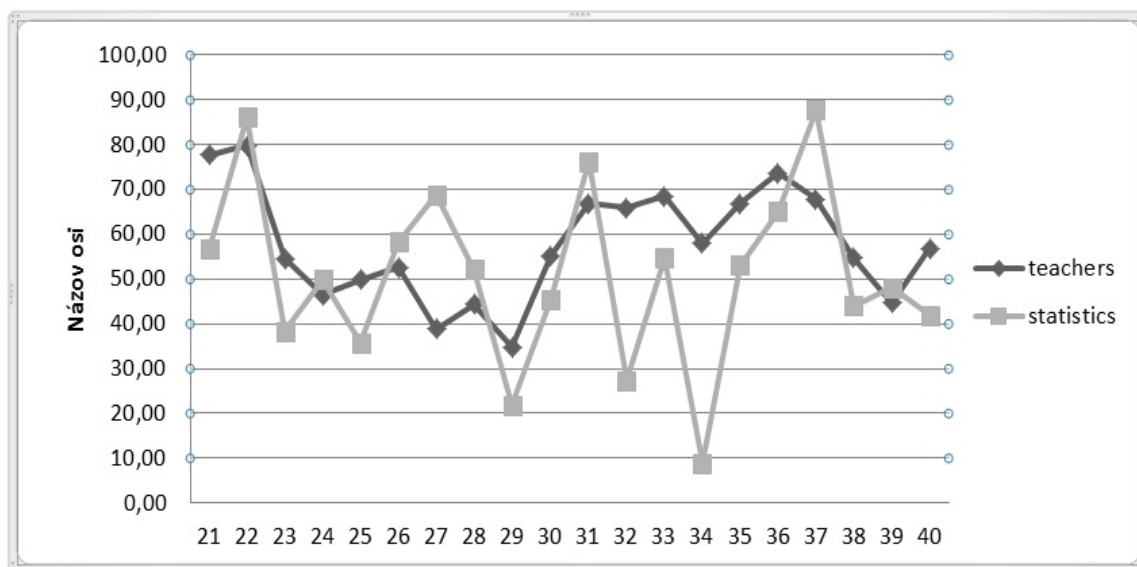


Figure 4 Teachers' judgements versus test scores in Language in Use (multiple-choice) 2014

The graph related to multiple-choice task 2014 shows that students had problems with selecting the correct words from the distractors. A good item is that one which discriminates strong students from weak ones (Bachman 2005), which was absolutely proved with problematic items 23 and 34. In both items only good students were able to respond correctly, while weaker students answered wrongly or left out the items. Lower performances are caused by the ministerial decision that all the students studying at the secondary grammar-schools were obliged to sit for a B2 test. Whereas these schools prepare students for their university studies in all academic disciplines, these students primarily focus

on their profiled subjects and English has become only a must for them, which influences their approach to its learning.

1.3 Reading Comprehension

Reading tasks measured comprehension in two closed-item tasks and one open-ended task. The teachers' expectations were consistently lower. Their students had been expected to find and comprehend main ideas (matching) and specific information (true/false + paragraph evidence and gap-filling), and some items seem to have been quite easy (63, 64, 65, 66, 68, 72, 73) or very easy for them (68, 76). On the other hand, standardized tests generally aim for a set of items with a moderate spread of difficulty values within the range of 30%-70%, which allows for maximum differentiation among individuals (Davies et al. 2002).

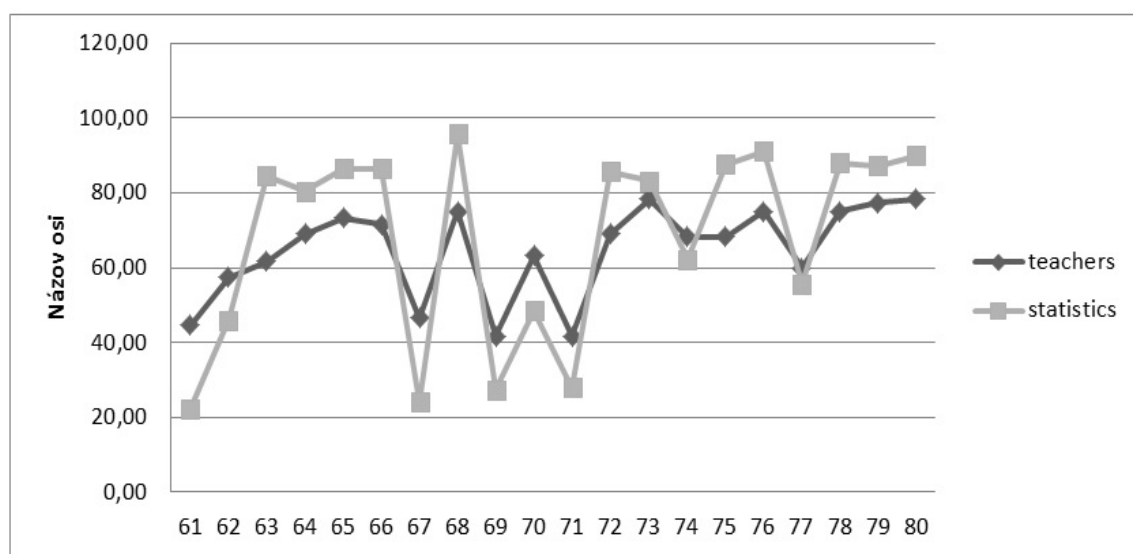


Figure 5 Teachers' judgements versus test scores in Reading 2008

Analyzing reading comprehension of 2014 school-leavers, teachers expected higher performances in Task 3 (74-80), in which students were expected to fill appropriate words in the gaps of statements related to the text provided. Influenced by previous years, teachers expected that words to be filled-in will be recognizable in the text provided. In 2014, students were expected to produce proper words due to their comprehension of relevant information.

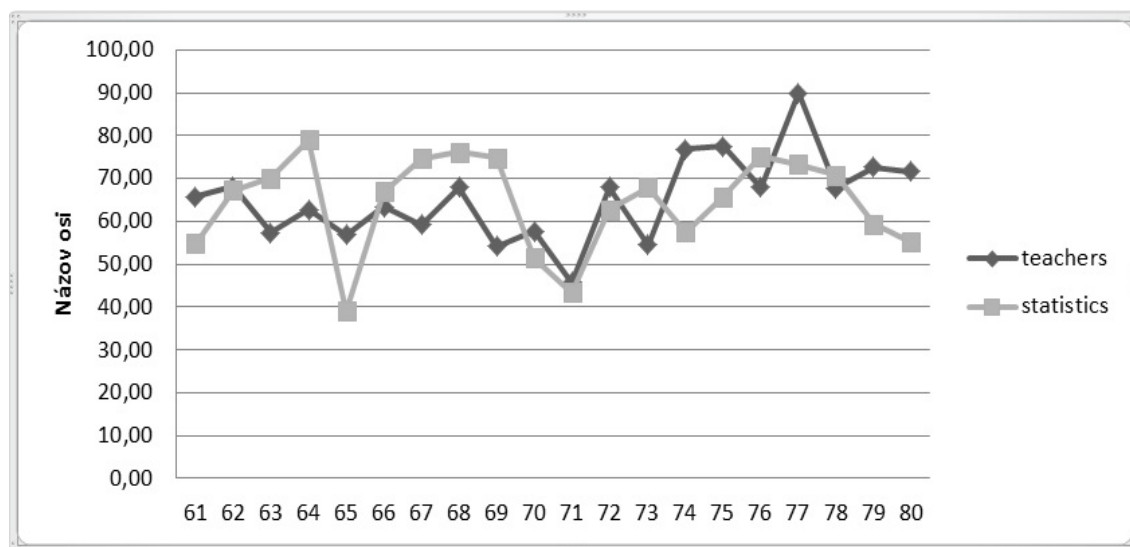


Figure 6 Teachers' judgements versus test scores in Reading 2014

We recognized that students did not make mistakes in finding proper information in the text, but in writing the expressions accurately, using appropriate morphological forms and syntactical functions. In their judgements, teachers were misled by the adequately comprehensible text and underestimated possible problems referring to grammatically appropriate and accurate answers. On the other hand, it is necessary to add that while in B1 tests students are expected to fill in the word from the text provided, B2 students should recognize meaning and then process that information into structurally and functionally accurate messages.

CONCLUSIONS

Recently, testing has gone through several steps of its development. The approach to validity has changed completely and the CEFR emphasizes that validity refers to demonstrating that what we claim is true. The CEFR's model of language use is called socio-cognitive as language is both an internalized set of competences and externalized set of social behaviors. If the focus is more on competence, then validity evidence will relate to cognitive skills, strategies and language knowledge that support inference about the potential ability for language use, for example a school-leaving examination (Manual for language test development and examining 2011). Confidence in a test is directly proportional to the amount of evidence that is gathered in support of its validity. The most important issue to be considered in validation of a test is that of its construct validity. This involves analyzing how test scores may be understood in relation to the theoretical framework underlying the construct the test is designed to measure (Davies et al. 2002).

The first workshops in 2008 and the correlation between the teachers' judgments and their scores (0.71) led to the second series of sessions in 2012. Since 2012 regular series of training sessions have been organized to enable teachers understand single descriptors calibrated for level B2. Comparing the 2008 judgments with those of 2014, we may conclude that teachers are more aware of the action-oriented approach presented by the CEFR, extending their teaching beyond the narrowly linguistic range by including pragmatic, sociolinguistic, intercultural, strategic and existential competences. Since conceptualization of language proficiency has shifted from an emphasis on linguistic form to one of communicative language ability and communicative language use, it is necessary to discuss the construction of reliable and valid language tests. In our country, it is still necessary to develop fair, objective and valid tests for measuring productive skills.

We compared item difficulty (defined as the proportion of test takers who answered the item correctly or the average score on the item) and judgements of teachers for a particular group of students and test items, to see how well matched they are. This comparison enabled us to see that some items which the teachers linked to the CEFR at B2 level were too difficult for some test-takers, but as we mentioned above, the ministerial decision that all the secondary grammar-school leavers should sit for a higher-level test could have an impact on the scores. While in 2008 teachers expected worse results in testing receptive skills, in 2014 their expectations were not one-sided. Teachers' judgements of multi-choice students' performances in 2008 were more consistent, while in 2014 they carefully judged every item.

To conclude, both pre-service and in-service teacher training should include language testing as one of their priorities. Teachers should be prepared how to assess their students' performances objectively and professionally. The exposure to the professionally-designed tests in classroom assessment should significantly reflect in the students' summative assessment.

REFERENCES

- Bachman, LF 1995, *Fundamental Considerations in Language Testing*, 3rd edn, Oxford University Press, Oxford.
- Bachman LF 2005, *Statistical Analyses for Language Assessment*, Cambridge University Press, Cambridge.
- Béřešová J 2011, 'The Impact of the Common European Framework of Reference on Teaching and Testing in Central and Eastern European Context', *Synergies Europe*, no. 6, pp. 177-190.
- Canale M & Swain M 1980, 'Theoretical bases of communicative approaches to second language teaching and testing', *Applied Linguistics*, no.1, pp. 1-47.
- Common European Framework of Reference for Languages: Learning, teaching and assessment* 2001, Cambridge University Press, Cambridge.
- Davies A, Brown A, Elder C, Hill K, Lumley T, McNamara T 2002, *Dictionary of language testing*, reprinted, Cambridge University Press, Cambridge.
- Hymes DH 1972, 'On communicative competence', in JB Pride & J Holmes (eds.), *Sociolinguistics*, Penguin, Harmondsworth.
- Littlewood W 1994, *Communicative Language Teaching*, 15th edn, Cambridge University Press, Cambridge.
- Manual for Language Test Development and Examining*, 2011, Council of Europe, Strasbourg.