

SUPPORTING OBJECTIVITY IN TESTING SPEAKING SKILLS

Jana Bérešová

Trnava University, Hornopotočná 23, Trnava 918 43, Slovakia

Abstract

Language teachers prefer testing grammar and vocabulary to testing productive skills as this assessment appears to be more subjective rather than objective. As far as productive skills are concerned, the testing techniques enable teachers to produce tests that meet the criteria of reliability, validity and objectivity. In our study, we introduced some variables and tested their effect on ensuring validity as well as objectivity. The data achieved in the workshops and comments of the teachers who participated in them will be presented and analysed with the aim to provide some recommendations for those who are involved in assessing speaking skills.

Keywords: *speaking skills, testing, objectivity, validity, authenticity, real-life tasks*

1. INTRODUCTION

Communicative testing has changed a traditional way of preparing tests concerning mostly isolated sentences in which particular vocabulary and grammatical patterns were the focus of getting information about learners' knowledge rather than their ability to use a target language. This way of testing still supports a traditional way of teaching in which the focus is put on language knowledge rather than communicative language competence. Teachers do not distinguish between exercises in which a particular grammatical pattern and a list of words are practised and tests that need to be valid, reliable and objective. They prefer testing grammar and vocabulary in similarly designed exercises that can give evidence what learners have learned without giving information whether learners can use a target language accurately and appropriately in real-life situations.

The Common European Framework of Reference for Languages: Learning, teaching, assessment (Council of Europe, 2001) presents four modes of communication: perception, production, interaction and mediation, supporting the concept of a learner-centred approach and that of an action-oriented approach. The focus on language knowledge has been shifted towards the learner's abilities to use a target language effectively, which requires learners to develop learning strategies that match their learning styles and become aware of learner autonomy concepts, enabling them to learn target languages for real-life purposes.

Despite the fact that teachers are aware of two approaches to teaching speaking (a direct or controlled and indirect or transfer), most teachers adopt the approach that is determined by a coursebook they use. Other limitations recognisable in many coursebooks account for problems with teaching certain speech acts, insufficient emphasis on teaching communication strategies and a lack of authenticity (Gogh and Burns, 2012). Therefore, a balance between both approaches (one focusing on language knowledge and another emphasising language use) should be encouraged in language teaching.

With regard to the CEFR (Council of Europe, 2001), the concept of profiling is based on a number of descriptor scales in order to enable language learners to identify that in real life they can develop differentiated profiles, taking into consideration their learning needs while learning and acquiring a target language. In most countries, the national curricula require language learners to achieve the same proficiency level in all language communicative skills. In Slovakia, B1 is expected to be achieved by learners studying at secondary technical schools and vocational schools, B2 at secondary grammar schools and C1 at bilingual secondary schools or bilingual sections of secondary schools. Therefore, in testing speaking, the schools themselves are obliged to prepare 25 examination papers that cover all the topics presented in the national curricula, based on a particular level. According to decree 318/2008 Coll. on study completion at secondary schools in Slovakia, each examination paper needs to be comprised of three tasks:

- a) a visual stimulus (describing visual stimuli while searching for relations between them and the task instructions), focusing on coherence, linearity in expressing the ideas and adequate language referring to a particular proficiency level in assessment
- b) a topic (adopting an attitude to the topic and justifying it), assessing learner's ability to express their opinions related to the topic and justify them
- c) a situation task (a dialogue divided by roles), assessing the appropriacy and accuracy of language in real-life situations (Béřešová, 2017).

Since this section of the school-leaving examination is not guaranteed by an external body, every school needs to follow the decree in order to prepare this part of target language examinations adequately, on a professional basis. This requires intensive training workshops for language teachers who are obliged to be aware of the expectations for each level, which needs being familiar with CEFR descriptors related to the three above-mentioned levels and to understand the aspects that enable them to ensure reliable, valid and objective speaking skills testing.

The country encourages both pre-service and in-service teacher education. Despite the fact that teachers of English are university educated people who have attended ELT courses focused on all the aspects covered by ELT during their master-degree studies, the level of quality of these academic courses is different due to the autonomous status of each university and its preferences. The number of those academicians who have been trained in test development, item writing and are professionally trained assessors or markers in language testing institutions is quite low. Their theoretical knowledge might be deep but their practical experience is limited as they produce only institutional tests, which are not properly analysed, supported by statistics such as item response theory, etc. From another point of view, university teachers have their own either institutional or professional preferences and conduct research in different areas of language teaching. Therefore, language testing need not be the focus of their academic courses.

2. TEACHER TRAINING

To make a gap between theory and practice narrower, the workshop for secondary school teachers was organised by an in-service teacher training centre in Trnava and teachers who expressed their willingness to actively participate in the training process were invited. The number of participating teachers accounted for 44, which means 73.33% of the total number of applicants. The admittance was based on teachers' engagement in item writing and assessing students' performances during the oral part of the school-leaving examination.

The workshop was held in March 2019, starting with a presentation which introduced the most important concepts such as reliability, validity and objectivity. In these workshops, great emphasis is put on the fact that every test needs to test what it is constructed for and intended to test as a number of teachers are not aware of the fact that many times they focus too much on testing grammar and vocabulary rather than testing the ability of learners to accomplish a task, the way learners are able to convey a message or organise a coherent text using proper cohesive devices while using effective language. Validity is usually presented as a central concept, the definition of which has been extended recently to encompass the way the tests are used and refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests (Council of Europe, 2011). According to Thornbury (2006), reliability can be an effect of the test design and the subjectivity of scoring. Objectivity is at risk when there is one examiner, or there are no scoring criteria. All the above-mentioned aspects require item writers and examiners to be trained in being aware of CEFR levels, writing tasks, scoring learners' performances against scoring criteria that are standardised, which was clearly and conceptually presented at the workshop, enabling teachers to comment on the issues and share their own experiences referring to item writing and assessing their students' performances.

2.1 A common understanding of the reference levels

Prior to training assessors and the standardisation of marking criteria, the participants of the workshop were expected to write tasks that will meet the criteria of a good task. Since the teachers represented different types of secondary schools in the country, it seemed to be necessary to familiarise them with descriptors concerning the three above-mentioned levels. The selection of scales that might provide teachers with a clear concept of both oral production and oral interaction was preceded by indicating reference levels in general scales, later completed by indicating them in more specific ones.

The first activity focused on indicating a level corresponding to the descriptor referring to salient features of CEFR levels A1 to C2, made up of six basic levels and three plus levels (Council of Europe, 2001: 33-36). As this general grid is related to spoken mode of language, teachers' success of indicating most levels correctly became motivational and encouraged them to continue in dealing with more scales. An overall oral production scale was complemented by the scale related to sustained monologue: putting a case while raising awareness of an overall oral interaction was intensified by the scale related to conversation (Council of Europe, 2018).

To produce something in spoken mode, language users usually can prepare their speech consciously, taking into account the effect of different styles, discourse structures or formulations. In real life, it is possible to rehearse the speech, which enables language users to monitor and repair linguistic and communicative problems that can be done while speaking in front of an audience. The scale presenting descriptors for overall spoken production does not contain plus levels and therefore teachers could focus on key words and their judgements are shown in Table 1.

Table 1. Teachers' judgements – familiarisation with an overall oral production scale

A1	A2	B1	B2	C1	C2
100%	79.54%	65.90%	59.09%	63.63%	72.72%

Analysing the results, it can be concluded that the levels important for local examinations were judged less consistently than those ones teachers are not engaged in while teaching and testing English. However, it is important to mention that B1, B2 and C1 levels contain several expressions that might have confused teachers such as *reasonably fluently* in B1, *a wide range of subjects* or *expanding and supporting ideas with subsidiary points and relevant examples* in B2 and *detailed descriptors* or *complex subjects, integrating sub-themes* at level C1. This misleading was caused by insufficient reading without taking a particular context into account and teachers placed a strong focus on isolated phrases.

However, the estimations concerning overall spoken interaction were even more inconsistent since A2, B1 and B2 levels are comprised of plus levels and detailed descriptions of a particular level were more complex, which resulted in less consistent judgements of the teachers. In Table 2 three main levels are depicted as judged by teachers.

Table 2. Teachers' judgements – familiarisation with an overall interaction scale

B1	B2	C1
61.35%	52.27%	70.45%

Teachers' judgements can be interpreted that if a band contains more descriptors which are comprised of similar wording such as *a wide range*, *fluently*, *spontaneously*, *can enter unprepared into conversations*, etc., teachers immediately matched the description with a particular level, and therefore they had not read the full text of a particular band. However, the least relevant judgements refer to

level B2 that is a very important level for students who would like to study at Slovak universities as they are expected to continue in language learning either in EAP or ESP academic courses.

The discussion that arose after indicating these overall scales encouraged teachers to undertake other activities based on scales referring to both spoken production and spoken interaction more quickly, which resulted in many wrongly indicated levels, mainly those referring to B2 as this level contains sets of descriptors concerning particular quality indicators related to mastery of the target language. This level supported by the descriptors such as *can engage in extended conversation on most general topics in a clearly participatory fashion, even in a noisy environment* was judged mostly as C1 due to the words such as *extended*, *most topics* and *noisy environment*.

2.2 Task writing

While constructing a task, it is important to be aware of what speaking skills and/or subskills will be tested. According to Goh and Burns (2012), core speaking skills entail pronunciation (producing the sounds of the target language at the segmental and supra-segmental levels), speech function (performing a precise communicative function or speech act), interaction management (regulating conversations and discussions during interactions) and discourse organisation (creating extended discourse in various genres according to socio-culturally appropriate conventions of language). In the CEFR (Council of Europe, 2001), speech functions are defined as micro-functions while discourse organisation is called macro-functions. Teachers have become aware of these facts and realised that one of the primary goals of testing speaking is that the ability of learners to use the language properly is to be measured objectively and assessing particular skills or subskills is to be reflected in the marking criteria.

As far as testing oral production is concerned, a good task consists of the rubric (clearly stated instructions, guiding students to accomplish a particular task), containing the topic, the length and what ensures that the task can be considered accomplished (macro- and micro-functions). The structured task is comprised of a number of issues students have to include in their speech as the topic-related points provide them with a sufficient amount of input, enabling them to produce enough material to be measured. The task comprising a topic, supported by relevant bullet points, enables students to elicit adequate scope to demonstrate their range and control of language.

With regard to testing oral interaction, the most appropriate seems to be playing roles in case students are provided with a sufficient amount of input in order to be able to develop the given topic. In this type of tasks, the students' performances of different communicative functions or speech acts can be measured as well as managing interactions and directing them. The performance of a student can be significantly influenced by an interlocutor and therefore it is important to decide who will be the person interacting with a student. In the local context, interlocutors are preferably teachers as they are considered to be methodologically prepared how to play their roles appropriately, in a balanced way, respecting all the issues concerning the level of proficiency, the difficulty of questions and statements, avoiding trick questions, etc.

To get a complete picture of a student performance, it seems to be appropriate to prepare different tasks used in one examination paper. The more tasks are given, the more complex assessment of learner's ability to use a foreign language in real-life situations can be achieved. These situations need to be realistic, yet based on learner's experience, matching his age and educational background. The aspect of authenticity can be ensured by the conditions under which the learner is expected to use a foreign language and therefore it is natural for him/her to use English as there is no other way of making himself/herself understood.

On the other hand, the task should not be knowledge-based. The assessor cannot base his/her rating on facts, data and depth of knowledge and learners cannot achieve higher scores due to their content knowledge. Some students might be more familiar with the topic and then it is difficult to distinguish whether their performances are good due to language competence or previous knowledge (Béřešová, 2017). If assessors prefer knowledge, learners learn the texts by heart and testing is based on testing their memory rather than their ability to use the language effectively. The successful achievement of

the task is also based on the ability of learners to express the above-mentioned micro-functions appropriately, accurately and naturally in the context provided.

Every task should be immediately completed by marking criteria to recognise whether the task enables teachers to assess learners' performances objectively, to develop the topic and perform according interaction management aspects, such as changing topics, clarifying meaning, etc. The marking criteria can help item writers (teachers) discover that being exposed to a lot of information encourages learners to use it without any effort to search for proper words or grammatical patterns to convey a message.

A good task should elicit communication strategies that competent language learners use in order to find alternative ways of expressing their meaning, keep their turn, or manage spoken discourse. Spoken production resembles a planned text that is more typical for written language and therefore some applied linguists call it a verbal essay as it has many salient features typical for written essays.

While producing their own tasks, teachers became aware of all the above-mentioned issues. They were working in pairs and each pair had to present their task in a group of six teachers. Teachers assess tasks against a list of criteria and prepared questions and comments for overall discussion and conclusions. The result of this session of the workshop was a set of tasks applicable immediately in their English classes. These examples of good practice are supposed to have an influential impact on item writing, impinging on 25 examination papers encompassing three tasks, exclusively prepared for school-leavers at three different levels of proficiency.

2.3 Assessor training

This part of training started with benchmarking of the calibrated spoken performances for three CEFR levels (B1, B2 and C1). The grid used for training was taken from the CEFR Table 3 (Council of Europe, 2001: 28-29) concerning qualitative features of spoken language since an extended one presented in the Companion Volume (Council of Europe, 2018) involves also a new aspect (phonology) that had been expected to be discussed so intensively that the workshop would never have finished. Other five aspects (range, fluency, accuracy, interaction and coherence) are frequently presented in in-service training workshops and teachers were familiar with the concept. The teachers were provided with illustrative sample performances, selecting those that were most relevant in terms of level and age-group. The teachers were expected to assign a level, first using a global scale, later focusing on the one comprising of the five above-mentioned aspects. All teachers' judgements were recorded and analysed. The discussion of the levels assigned in relation to the descriptors revealed that teachers' judgements were influenced by their own teaching experience and they did not read descriptors properly. Then the levels assigned to the performances that had been calibrated by a group of experts were announced and the official evaluation was presented and linked to the descriptors.

The candidates' performances, officially calibrated by a number of experts from a variety of European countries, were taken from the DVD that is a product of a seminar organised in 2008 by the CIEP (Centre international d'études pédagogiques) in France with the assistance of the Language Policy Division of the Council of Europe. Teachers' inconsistency can be proved by the following example, in which two B1 officially calibrated performances were judged by participating teachers as follows: B1+ (43.18%) for a girl's performance and A2+ (40.91%) for a boy's performance.

Table 3. Teachers' judgements of B1 performances

	A2	A2+	B1	B1+	B2
Student A		6 (13.64%)	17 (38.64%)	19 (43.18%)	2 (4.54%)
Student B	17 (38.64%)	18 (40.91%)	7 (15.91%)	2 (4.54%)	

Table 3 shows that while a girl's performance was estimated to be at a higher level of level B, a boy's performance was viewed as an A-level performance. In addition to the judgement of the first performances, officially calibrated at B1 level for both speakers, an intense discussion occurred as teachers did not recognise that recorded performances had not been prepared in advance as it is done in the school-leaving examination when students are given 15-20 minutes to read the tasks and prepare their responses, using dictionaries and other supplementary materials. In their initial impression judgements, spoken performances were estimated at lower levels: B1 for a girl and A2 for a boy. The boy's immediate response to the task relates to B1 descriptors.

However, it is necessary to emphasise that teachers did not employ the grid properly since they compared the presented performances despite the clearly stated instructions. Moreover, they did not take into consideration the role of marking in assessing language communicative skills due to which different marks are used to assess particular performances that can still represent the performance standardised for a particular level.

When assessing students' performances was undertaken, the teachers were given two marking scales: holistic and analytic. Despite the fact that the holistic scale was considered much closer to classroom assessment as it focuses on the general impression gained from the language produced, the judgements of teachers were inconsistent and they admitted their subjective interpretation of particular descriptors linked to learners' performances. One band of a holistic scale, used in the workshop, is presented in Appendix A.

In general, holistic scoring reliability is considered lower due to bands written as a whole, while construct validity is viewed the same as in case of analytic scoring since different aspects of speaking are developing in the same rate while captured in a single score. According to Weigle (2002), holistic scoring needs time-consuming training and has the first impression effect. While assessing holistically is a more natural process in real life, analytic assessment provides diagnostic information. Reliability of an analytic scale is higher than holistic due to specific criteria and different aspects of speaking ability are considered to be developing at different rates.

According to the workshop participants, the analytic scale breaks the language produced down into a number of points for various aspects of language use and evokes the impression of strictness. However, they preferred a list of specific descriptors that aided them to focus on their students' performances (Appendix B). While being trained, teachers confirmed that it was easier for them to better understand particular descriptors related to different areas of speaking skills than relate students' performances to one band.

The local standardised performances were shown to teachers and they were expected to mark the individual students' performances. Their judgements were recorded in the grid based on the marking criteria and teachers realised that their assessments concerning grammar and vocabulary were more consistent than when assessing the aspects of task achievement or organisation and cohesion. Over 56% of teachers prioritised facts and knowledge over students' ability to express their ideas effectively. Their explanation was based on examining other subjects in the school-leaving examination in the country, where a knowledge-based approach is adopted. Fewer teachers (over 35%) admitted that it had been difficult for them to forget better learners' performances and not to compare their abilities with those of weaker learners, using exclusively the descriptors related to marking.

3. CONCLUSIONS

Despite the fact that the language curricula and an external part of the school-leaving examination are related to the CEFR, teaching English in primary and secondary schools are significantly influenced by intensive grammar and vocabulary teaching. The way of teaching is reflected in the way of testing and therefore most classroom tests contain the measurement of accurate language use (a large number of grammatical and lexical items) in isolated sentences or short texts at the expense of assessing receptive and productive skills.

When testing speaking, most teachers test memory rather than language competence, which many students consider easier than using a target language naturally and prefer this way of assessment in order to get better marks. Therefore, the goal of the workshop organised for teachers who are engaged in an internal part of the school-leaving examination in English was to introduce the principles and techniques referring to testing speaking in order to guide teachers in preparing their own tasks, taking into account all the aspects that make their tests valid and reliable. To enhance teachers' confidence in producing good tasks and measuring their students' oral performances as objectively as possible, this intensive workshop was held to train mostly those teachers who had unwillingly become item writers and assessors in their local contexts due to legislation. The aim was to assist teachers to become able to assess their students' speaking abilities, respecting both language knowledge and language use.

Having been familiarised with CEFR levels, teachers were given recommendations for writing good tasks and undertook training related to assigning a reference level to tasks and performances in the relevant skills. The workshop confirmed that teachers did not distinguish between rating and scoring as the former deals with assessing learners' performances against specified criteria, while scoring contains determining what mark or grade the learners should get, based on the assessment against the marking criteria. Comparing both holistic and analytic scales, teachers came to conclusions that focusing on the discrete aspects can lead to undesirable strictness, and therefore an overall impression based on holistic scoring should be included to adjust the final decision.

The views of some applied linguists that each task should be completed by a task-specific marking scale as they differ to a particular extent, controlled by input, were introduced and supported by several examples. Despite being exposed to a variety of approaches to testing speaking skills, teachers accepted the importance of meeting all the requirements which enable assessors to provide learners with valid and reliable scoring. They agreed upon the fact that intensive training they had undertaken enabled them to internalise the concepts and ideas related to the aspect of objectivity.

Objectivity in testing speaking was viewed achievable through a good task design that enables students to provide enough evidence about their language competence. The concept of authenticity concerning a real-life task and conditions under which the task is accomplished was clearly introduced and teachers grasped the significance of the clearly formulated rubric as there is no room for confusion and need for clarification, as both are likely to create anxiety in the students and impair their performances and affect the reliability and validity of the test.

ACKNOWLEDGEMENT

The author would like to express gratitude to the Slovak teachers of English, who had voluntarily participated in this research study, for their open-minded approach and the willingness to invest their time in being members of the panel.

REFERENCES

- ALTE Manual for Language Test Development and Examining*, 2011, Council of Europe, Strasbourg.
- Bérešová J 2017, *The Impact of the CEFR on Language Examinations in Local Contexts*, Peter Lang GmH, Frankfurt am Main.
- Common European Framework of Reference for Languages: Learning, teaching, assessment* 2001, Council of Europe, Strasbourg and Cambridge University Press, Cambridge.
- Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion Volume with New Descriptors*, 2018, Council of Europe, Strasbourg.
- Goh CM & Burns A 2012, *Teaching Speaking. A Holistic Approach*, Cambridge University Press, Cambridge.
- Thornbury S 2006, *An A-Z of ELT. A Dictionary of Terms and Concepts*, Macmillan, Oxford.

APPENDIX A - An example of a holistic scale band

Band 5 Full realisation of task set shown by:

- coverage of points required with evidence of original output
- wide range of structure and vocabulary demonstrating control of language
- clear organization with a variety of linking devices
- presentation and register wholly appropriate to purpose and audience throughout

Overall: a very positive effect on the target reader (UCLES 1997).

APPENDIX B - An analytic scale

Content

- 5 – relevant, fully developed and comprehensive
- 4 - relevant, well-developed, some information missing
- 3 - mostly relevant, adequately developed, a few pieces of information missing
- 2 - limited relevance, most of information missing
- 1 - very limited relevance, hardly any information included
- 0 - totally inadequate content

Organisation

- 5 - coherent text, a wide range of linking devices, proper paragraph development
- 4 - clear, well-structured, good use linking devices, some hesitation in paragraph development
- 3 - clear relationship between main ideas, limited use of linking words, paragraph development restricted to main ideas
- 2 - linear sequence of ideas, limited number of linking words, paragraphing inappropriate
- 1 - ideas confused and disconnected, basic linking words, no paragraphs
- 0 - no apparent organisation of content

Grammar

- 5 - accurate use of grammatical structures appropriate to the task
- 4 - a few minor mistakes only, mostly complex sentences
- 3 - a few major mistakes, but a lot of minor errors, mostly simple sentences
- 2 - major mistakes, communication difficult
- 1 - numerous serious mistakes, communication very difficult
- 0 - many serious mistakes, almost no communication

Vocabulary

- 5 - a wide range of vocabulary appropriate to the task
- 4 - good use of vocabulary appropriate to the task
- 3 - vocabulary sometimes restricted, do not interfere significantly with communication
- 2 - restricted vocabulary, communication quite difficult
- 1 - very restricted vocabulary, communication is very limited
- 0 - vocabulary inappropriate to the task, communication very difficult (Bérešová, 2017: 146).