Data elicitation through language testing: Challenges of test design

Monika Černá
University of Pardubice, Czech Republic

Abstract: *The paper discusses the challenges of test design in the context of a research project focusing on the analysis of tertiary students' spoken production in English. One of the project aims is to create a corpus of learner-spoken English. The participants in the study are Czech first-year students in English language teacher education study programmes at three universities. In order to elicit samples of the students' oral production, a test of speaking, including a pronunciation subtest,was designed with respect to the research aims and objectives and in accordance with the current trends in the field. The challenges faced by the research team may be divided into three groups – those pertinent to the construction phase of the research instrument, the pilot phase, and the data-collection phase. The paper discusses how the team responded to the perceived challenges.The process of test designing was informed by relevant literature (e.g. Bachman, 1990, Hughes, 2003, Luoma, 2004);the team strove to achieve the highest possible level of 'test usefulness', i.e. test qualities including reliability, construct validity, authenticity, interactiveness, impact and practicality, as introduced by Bachman and Palmer (2009). Consequently, the decisions regarding the abilities that the candidates, prospective English teachers, should demonstrate, as well as the decisions about the test format (elicitation techniques, number of tasks, etc.), were made with respect to the proposed 'test usefulness'. The pilot phase confirmed the usability of the tool to elicit the required data, but also necessitated a few content- and procedure-related modifications. They reflected the results of the analysis of the performances recorded during the trial testing as well as the analysis of feedback questionnaires.Having revised the test, the researchers then implemented it in the data-collection processin three universities in the Czech Republic. The number of studentstested was 176.*

Keywords: *Diagnostic test, speaking, English, tertiary students, research*

Article History:
Submitted: 15.04.2014.
Accepted: 08.11.2014.

DOI Number:
10.14706/JFLTAL152115

Introduction

Data elicitation is inherent to any research, not excluding second language research. Depending on the aims of a study, various data elicitation techniques are used. Gass and Mackey (2011) place techniques on a continuum ranging from naturalistic data to prompted-production data and prompted-response data. Regarding the investigations of input and interaction, there has been a gradual move away from studying those aspects in natural settings (Ellis, 2008). Researchers often rely on clinical elicitation, i.e. prompted production and prompted response, through which samples of learner language are obtained. Gass and Mackey (2011) suggested specific examples of diverse elicitation techniques. Additionally, language tests may also be utilised to elicit data for a variety of research purposes such as 'research into the language ability itself, including the effects of different test taker characteristics on language test performance' (Bachman & Palmer, 2009, p. 99). This matches the focus of the research project[1] designed to investigate the influence of the Czech students' mother tongue on their communicative competence in spoken English in relation to the students' individual learning histories (Černá, 2013). The project is discussed in the paper with an emphasis on the challenges of test designing within its context.

Achieving a high level of communicative competence in the target language has been the core of the language teacher's expertise. Therefore, the project of a diagnostic nature has been targeted at Czech students' spoken production in English on the onset of university teacher education. The findings of the project will provide insights into the processes and outcomes of learning English as a foreign language in the Czech Republic, which may be found beneficial by educational institutions operating at all levels from the pre-primary to upper-secondary, namely for the purpose of curriculum design. The results will also function as feedback for Czech authors of English textbooks and as a basis for the design of new research-based teaching materials. Most importantly, the outcomes will be utilised in teacher education, research-based evidence will allow for the development of methodologies leading to a sound content knowledge base of trainee teachers. The project started in 2013; actions in the first year were centred on data-elicitation tools.

Challenges of test design

Testing or formal assessment in general is a complex and challenging matter; it is even more so with testing spoken language. Davis (2009) attributes the complexity to

[1] A three-year project, *Aspects of English Language Acquisition of Czech Students on the Onset of Teacher Education*, has been supported by the Czech Science Foundation (GA ČR 13-25982S).

the interaction of different factors and their influence on a final score. It is far beyond the scope of the paper to attempt to reiterate all the relevant factors, therefore, the major challenges perceived in the particular testing situation will be in focus.

Designing a test for research purposes

The priority of research is obvious in the testing situation explored in the paper. The team was supposed to construct a diagnostic test of spoken language that would elicit samples of learner language for the subsequent investigation of its variability. A set of structures to be included in the analysis has been identified on the basis of the following criteria(a) relevance of particular features to the grammar of spoken language, and (b) potential negative transfer from the mother tongue. Selected syntactic and discourse features comprise word order deviations (both incorrect and systemic caused by the nature of conversation), distribution of verbs typical of spoken discourse (functioning as discourse markers, main clauses or comment clauses), structures with non-finite verb complementation and the use of vagueness hedges (Ježková, 2012). Regarding pronunciation, the segmental and supra-segmental features of interest include the following: the front open vowel 'ash', the weak central mid vowel 'schwa', the voiced and voiceless dental fricatives, the labiovelar approximant /w/, the velar nasal, the pronunciation of word-final voiced consonants and non-initial primary word stress(Nádraská, under review). Apart from data elicitation the test administration is expected to impact on the students in the first year of the English major bachelor study programmes at three universities in the Czech Republic involved in the project. Being a diagnostic tool, the test should uncover the students' strengths and weaknesses in performing oral communication tasks. The diagnosis on entry to the programmes may lead to possible adjustments of syllabus objectives of relevant courses, e.g. language development courses, phonetics, phonology and syntax. Furthermore, the performance on the test is likely to influence the setting of the students' autonomous language development goals.

Considering the features of the particular testing situation, the research team aimed to achieve the highest possible level of 'test usefulness', which was proposed by Bachman and Palmer as 'a function of several different qualities, all of which contribute in unique but interrelated ways to the overall usefulness of a given test' (2009, p. 18). These include reliability, construct validity, authenticity, interactiveness, impact, and practicality; out of the listed qualities, authenticity will be at the centre of attention. Bachman and Palmer consider authenticity a critical quality of language tests (2009, p. 23) and define it 'as the degree of correspondence of the characteristics of a given language test task to the features of a TLU [target language use]task.' (ibid.). In order to achieve the highest possible level of authenticity, the team attempted to design a test that would be relevant to the target language use domain, i.e. that of teaching English as a foreign language. Reflecting

the variety of the target language-use tasks, it was desirable to design test tasks capable of eliciting samples of both monologic production and of student-student spoken interaction. Therefore, the paired format proved a necessity.

The researchers were aware of the advantages and caveats of the paired format reported in the papers on testing speaking (e.g. Galacsi, 2010). In discussing various task types O'Sullivan (2012) concluded that there were issues for less outgoing studentsin relation to interactive tasks. However, there seem to be other assets of the paired format that are worth considering. Galacsi (2010) enumerates studies that support the finding that oral paired tasks were more symmetrical in the interaction possibilities they created. Brooks (2009) reports that subjects in her study performed better in the paired format than they did in the individual format. While the latter tended to result in asymmetrical discourse, a variety of interactive features was distributed in a more balanced way in the former paired format. With the research aims in mind, those findings provided a substantial argument for involving the paired format. Nevertheless, there were other questions to answer, namely those related to the ways of pairing the test-takers. Two factors, reflecting the project aims, will be mentioned: the influence of interlocutor proficiency and learner acquaintanceship. The subjects in the study are students on entry to the tertiary education, i.e. the diagnosis is scheduled as soon as possible after the beginning of the academic year to prevent the impact of university education to contaminate the data. Before the diagnosis there is virtually no possibility for the researchers to learn either about the students' proficiency in English or about their social relationships in the newly constituted groups. Although there is some research evidence that subjects achieve higher scores when working with a friend (O'Sullivan, 2009), the acquaintanceship effect was ignored in this particular testing situation. The examinees could choose a partner, but it was based on availability rather than personal preference;however,thepotential existence of some interpersonal relationships cannot be excluded. Regarding a variety of proficiency levels, it was considered in the light of the study by Davis (2009) in which he investigates the effects of the proficiency level of an examinee's partner in a paired oral test. Davis concludes that the level of proficiency has little influence on scores, but in some cases the pairing type appears to influence language quantity or interaction characteristics (ibid.). In the context of the research project, a potential decrease of language quantity or eliciting a type of response other than expected would have detrimental impact on the obtained data. In order to prevent this, two information-exchange tasks with precisely defined roles were included together with an informal discussion. Whether the three interactive tasks provide space for each test-taker to produce the expected response should be verified in the pilot phase of the test construction process. Reflecting the research aims and with reference to relevant resources (e.g. Bachman, 1990, Hewings, 2004, Hughes, 2003, Luoma, 2004) and preliminary studies (Černá, Urbanová, &Vít, 2010,

Ježková, 2012), the team constructed a diagnostic speaking test with a pronunciation subtest. The table below presents selected test tasks' characteristics based on the framework proposed by Bachman and Palmer (2009).

		TYPE OF TASK	FORMAT	INPUT	EXPECTED RESPONSE	INPUT – RESPONSE RELATIONSHIP
SPEAKING	Introduction	Warm-up	Individual	Aural, Target language*, Language input: sentences, prompt = open-ended questions, Unspeeded*, Live*	Oral*, Target language*, Limited production response, Unspeeded*, Live*	Reciprocal, Narrow scope, Indirect
	Task 1	Sustained monologue	Individual	Aural, Language input: sentences, prompt = open-ended questions,	Extensive production response, Individual long turn	Non-reciprocal, Narrow scope, Indirect
	Task 2	Information transfer (asking/ giving detailed information about events, processes; telling what to do)	Paired	Visual, Language input: words, phrases, sentences, prompt = task sheet, Non-language input: pictures	Co-constructed, extensive production response, Transactional and interactional language	Reciprocal, Broad scope, Direct
	Task 3	Information transfer (see Task 2)	see Task 2	see Task 2	see Task 2	see Task 2

	Task 4	Informal discussion	Paired	Visual, Language input: phrases, sentences, prompt = issue to discuss, clues given	Co-constructed, extensive production response, Transactional and interactional language	Reciprocal, Narrow scope, Indirect
PRONUNCIATION	Task 5	Reading aloud: text	Individual	Visual, Language input: extended discourse, prompt = text (152 words)	Extensive production response	Non-reciprocal, Broad scope, Direct
	Task 6	Reading aloud: word list	Individual	Visual, Language input: words, prompt = wordlist (27 words)	Limited production response	Non-reciprocal, Broad scope, Direct

*The characteristics that remain the same are not repeated for each task.

Not only the test characteristics but also the topical content of a test plays an important role with respect to its authenticity. In this particular testing situation, the researchers explored a range of topics that would be appropriate to the test-takers with the following characteristics: young adults on the onset of their university teacher education, native speakers of Czech, the level of communicative competence approximately B2 according to the Common European Framework (Council of Europe, 2001). Topical knowledge was deliberately excluded not to favour certain test takers; all the information necessary to complete the tasks was prompted in the input or personal experience was called for. The topics were carefully considered to avoid those that might be perceived as sensitive by the test-takers; for example, the following topics were finally involved in the test: experience with learning English, renting a flat, student mobility, part-time jobs, the role of social networks in one's life, healthy eating, and plagiarism. Although overreaction to any of the topics was not expected, topic relevance was also examined in the pilot phase.

Trial testing

Challenges of the pilot phase were manifold. Since the respondents were recruited on a voluntary basis, the main challenge was to attain a sufficient number of cooperating students with such a set of characteristics that would be close to those of the prospective cohort. Furthermore, the implementation of the trial version of the test

was seriously constrained by the schedule of the academic year. In spite of all the problems the test was piloted with a group of first-year students. The performances were recorded and four of them were selected for a detailed analysis (fivewomen, three men). All the participants in the pilot study completed a feedback questionnaire after the performance and were invited to discuss any aspects of the performance with the researchers. The questionnaires and outcomes from the discussions were investigated too. Consistent with the objectives of this paper, only selected outcomes of the analyses will be presented, i.e. those focusing on the quantity of language produced by individual students inthe interactive tasks and topic relevance.

The analysis of the recorded performances was primarily targeted at test-takers' participation in the interactive tasks. Individual students varied in the total time spent onthetasks, and individual students alsospent a different amount of time on each of the tasks (see Chart 1 below). However, to judge the language quantity, the number of words is used as a criterion. When considered in relation to time, the difference between student 1 and student 2 (S1 – S2) in pair 1 has slightly diminished, the variation in pair 2 (S3 – S4) remained roughly the same but the differences in pairs 3 (S5 – S6) and 4 (S7 – S8) have magnified considerably (see Chart 2). Overall, the testees' personal attributes, along with thetopic and task characteristics, may account for the variation. Given that tasks 1 and 2 are in principle the same, the difference may be attributed to the topic (e.g. S8). Task 3 is of a dissimilar nature; therefore, it is uneasy to uncover the reasons for variation. Hypothetically, they may be linked to the task characteristics or topical content.

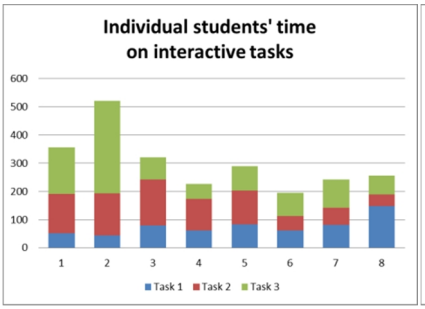

Chart 1

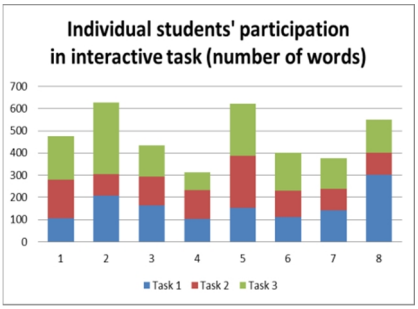

Chart 2

To find out the subjects' opinions about the topics, feedback questionnaires were analysed. Topics of individual tasks were evaluated positively; the respondents characterised them as relevant, useful and adequate to their life experience. Several topics initiated a certain level of emotional arousal; however, it concerned only positive emotions and the students appreciated it. No sensitive or inadequate topics were identified. Critical comments concerned the topic of plagiarism; interestingly, it was marked as irrelevant by pair 3. Obviously, none of the students had problems

discussing it (S5 – S6, Task 3). Observed quantitative differences in language production may be attributed to topic-related personal preferences.

Regarding the task characteristics, a few respondents pointed out that there was too much information on a task sheet. Consequently, all the sets of task sheets were revised in terms of language, informational relevance, and layout before the data-collection phase.

Test administration

Finally, the challenges experienced in the data collection phase should be mentioned. Since the research was conducted in three institutions located in diverse regions in the Czech Republic, it was demanding to prepare a schedule suitable for the participants as well as for the research team. Standardising the process of the test administrationwas another issue. It concerned not only testing conditions in the three institutions but also procedural aspects of the test. The total number of recorded students was 176. Threeacademics were involved in the data collection. As implied by the charts above, they occasionally failed to maintain internal consistency of time management. In situations when the discussion was evolving smoothly, the researchers provided discussants with unlimited time to finish the task.

Conclusion

The paper has deliberated the process of test construction in the context of a research project. While test has proved a valid technique of data elicitation, at the same time there seem to emerge certain tensions or potential conflicts. Most importantly, there exist conflicting needs of the research and those of the cooperating institutions. For example, the project required testing the students who meet the criteria to be included in the research sample; however, the test could fulfil its diagnostic function only if all the students were involved in the assessment. Thus there was an increased workload on the part of the researchers on the one hand but a positive impact on the studentson the other hand. Obviously, the research benefits the cooperating institutions, but at the same time interferes with their established procedures. Furthermore, the schedule of the research project is not necessarily in harmony with the academic-year schedule and time becomes a real issue.Lastly, there is an internal conflict of the two identities of the same person – that of a test designer and that of a researcher. The conflict is manifested in making decisions throughout the entire process of test construction. Apparently, project aims are prioritised and the decisions tend to be ‘research-friendly’.

References

Bachman, L. F. (1990).*Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2009). *Language Testing in Practice*. Oxford: Oxford University Press.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*,*26* (3), 341 – 366. doi: 10.1177/0265532209104666

Černá, M. (2013). Aspekty osvojování si anglického jazyka u českých studentů při vstupu do přípravného vzdělávání učitelů.In *Sborník příspěvků 21. výroční konference ČAPV Efektivita vzdělávání v proměnách společnosti*.Ústí nad Labem: UJEP.

Černá, M., Urbanová, Z., & Vít, M. (2010). Pronunciation awareness of students in English language study programme. In S. Válková (Ed.),*New trends in educating future teachers of English VI.* Olomouc: Palacký University.

Council of Europe (2001).*Common European Framework of Reference for Languages*.
Cambridge: Cambridge University Press.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*,*26* (3), 367 – 396. doi: 10.1177/0265532209104667

Ellis, R. (2008). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Galacsi, E. D. (2010). Paired Speaking Tests: An Approach Grounded in Theory and Practice. In J. Mader, &Z. Ürkün (Eds.), Recent Approaches to Teaching and Assessing Speaking. IATEFL TEA SIG Conference proceedings. Canterbury: UK: IATEFL Publications.

Gass, S. M., &Mackey, A. (2011). *Data Elicitation for Second and Foreign Language Research*. New York: Routledge.

Hewings, M. (2004). *Pronunciation Practice Activities*. Cambridge: Cambridge University Press.

Hughes, A. (2003).*Testing for Language Teachers*. 2nd edition. Cambridge: Cambridge University Press.

Ježková, Š. (2012). Syntactic Deviations and Mistakes in Learners' Spoken English Language. In R. Trušník, K. Nemčoková, &G. J. Bell(Eds.), *Theories and Practices – Zlín Proceedings in Humanities, Volume 3*(pp. 223-232). Zlín: Tomáš Baťa University.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

Nádraská, Z. (under review) A Pre-Diagnostic Stage of a Research Project: Selected Features of University Students' Pronunciation of English.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*,*19* (3), 277 – 295. doi: 10.1191/0265532202It205oa

O'Sullivan, B. (2012). Assessing Speaking. In C.A. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge Guide to Second Language Assessment*. New York: Cambridge University Press.