doi:10.5128/ERYa9.01

COMPARING SPEAKING SITUATIONS IN THREE DIFFERENT LANGUAGE TESTS

**Sari Ahola, Tiina Lammervo, Reeta Neittaanmäki,
Sari Ohranen, Henna Tossavainen**

Abstract. The purpose of this study was to compare learner outcome when five similar speaking situations were offered in the speaking subtest in Finnish, Swedish and English intermediate level tests in the Finnish National Certificates language testing system. The overall aim was to investigate the tasks by comparing learner outcomes across the three tests and to seek explanations for the outcome in the learners' self-reported demographic, language, educational and professional background.

The speaking situations were selected from the NC item bank meaning that they have undergone the Item Response Theory based analyses which indicate that the tasks function well in all tests. More information was needed to discover possible connections between performance and background information. Performance data and background factors were analysed using descriptive frequency and percent distribution. Cross-tabulation was used to analyse connections between variables. The tasks are discussed in terms of domains and language functions and their connection with the test takers' background information. Results indicate that though there is some variation in the learner outcomes across languages, situational tasks can be used for different languages.

Keywords: second and foreign language testing, spoken language, validity, Finnish, Swedish, English tests

1. Introduction

This article discusses the results of a study which compared learner[1] outcomes in a language test situation when the same test tasks were used in three different test languages. In the administration of the autumn 2011 test for the National Certificates (NC) in Finland, five situational tasks were offered in the speaking subtests of the

[1] Learner is used as a general term for language learners; test taker and candidate refer to people who take the NC examination.

English, Finnish and Swedish intermediate level tests. Additional data for Swedish was collected in spring 2012. The overall aim was to investigate differences and similarities in the performance of test takers across the three language tests and, furthermore, to seek explanations for the outcome in the candidates' demographic, language, educational and professional background as reported by them. This all relates to the general idea of validity in the use of situational test tasks in testing speaking skills in three different languages.

The NC is a second- and foreign-language proficiency test for adults in which the levels of language proficiency are linked with the Common European Framework of Reference for Languages (CEFR) 2001. The intermediate level tests measure language proficiency at independent language user levels B1 and B2. The test system as a whole is based on Finnish legislation (Act 964/2004; Decrees 1163/2004 and 1109/2011) and is independent of any syllabus or curriculum.

There are nine languages in the test system; the three languages in this study can be regarded as high-stakes. The test with the largest number of candidates, the Finnish test, is often taken by immigrants to demonstrate their language skills for immigration purposes. Finland has two official languages, Finnish (91% of population) and Swedish (5.4%), so the Swedish test can also be taken for the same immigration purpose. Finland has a long tradition of foreign language learning and currently English is the most popular foreigh language studied in schools. The English test is commonly taken by Finns who wish to work for the Finnish Defence Forces in international military cooperation roles. Furthermore, the results of all three tests are used as proof of language skills for professional purposes. The Swedish test is taken, for instance, by nurses and other health care professionals working in the Swedish-speaking areas in Finland.

The Finnish tests are taken by candidates with diverse language backgrounds (approximately 170 first languages overall) while the English tests are typically taken by Finns with Finnish or Swedish as their first language (L1). The Swedish test is taken by candidates from both diverse and Finnish language backgrounds.

The Framework of the Finnish National Certificates (2002, 2011) and the NC test specifications, which apply to all tests in the nine languages, provide the guidelines for item writing, test compilation and test development, with information about what constitutes communicative language proficiency and hence also the communicative language ability to speak in the foreign language. The test framework in the NC is based on the models of communicative competence offered in Canale and Swain (1980), and further in Bachman (1990) and Bachman and Palmer (1996), and in the framework offered in the CEFR 2001. Thus the test is designed to evaluate everyday language used for communication purposes rather than focusing on grammatical structures and vocabulary. The test specifications offer a defined set of topic areas and language functions for the purposes of item writing.

The test measures language ability in the subtests of writing, speaking, listening and reading. The tasks discussed in this article belong to the speaking subtest and comprise one of a total of four tasks in the subtest which are carried out in the language laboratory in all three tests. It is stated in the test specifications that the test tasks should be communicative and functional. It is also stated in the specifications that the tasks should take into consideration the interactional view of authenticity, which means that the tasks should consider the interaction between the test taker's language ability – level of proficiency and how to engage the test taker in the performance – and such situational task characteristics as the features of context and the test task (see ALTE Manual 2011: 12, Douglas 2000: 18).

The overall aim of the current study is to find out if the same test tasks are equally valid in all three subtests of speaking. Validity here refers to context validity since it accounts better for the social dimension of language use than the earlier concept of content validity (Weir 2005: 19). Weir (2005: 20) also states that though there are problems in ensuring that we actually follow the specifications and domains we promise to follow, and that there are problems in the operationalisation of every-day language use in a test situation, we should make attempts to ensure context validity.

Comparing test performance with background information (biodata) functions as a method of construct validation. The intention is to detect bias in the test for or against groups of test takers defined by their particular data (age, gender, education etc.) (Alderson et al. 1995: 185). The current study looks at selected background information to seek possible explanations for the differences in test taker performance.

2. Data and methods

The subtest of speaking consisted of four sections in which the candidates were asked, first, to describe an event, then take part in simulated conversations, react to situations and give a short speech. This study focuses on the candidates' reactions to situations which here are regarded as tasks and in the statistical analyses as items. The five situations represented different levels of formality and domains (personal–work), and also served different language purposes (as per NC item writing specifications). The following functions were required from the five test tasks:

1. The informative function of describing one's hobby and the interactional function of tempting a friend to join in.
2. The informative function of describing how the test takers learn new words and the possible methods that help them remember these words.
3. The emotive function of expressing feelings to a colleague in an upsetting situation (the test takers were asked to come up with the situation)
4. The interactional function of congratulating a friend on becoming a parent.
5. The informative function of describing the qualities of a good boss.

The speaking subtest was taken in a language laboratory where candidates' production was recorded. At the start of each section candidates had time to familiarise themselves with the instructions. A preparation time of 15–20 seconds was reserved for the candidates to prepare before each task/item (read the prompt text) and they had 30 seconds to complete their speaking. Candidates used the test booklet with instructions and task descriptions. Prompts advising when to start speaking came from the master recording.

The language of the test task instructions was not the same in all three tests. In the English test the instructions were in Finnish or Swedish, in effect, the test takers' chosen language of administration. In the Swedish and Finnish tests the instructions were in the target language. However, in all three languages the prerecorded prompt heard from the master tape giving the instructions of what to do (e.g. 'start speaking now') were in the target language.

The candidates' recorded speech, their test performance, was assessed using the NC criteria for speaking, calibrated and empirically linked with the CEFR. All NC tests are assessed by trained, registered raters[2]. All tests are assessed in centralised assessment sessions organised at the University of Jyväskylä. Every assessment meeting starts with a two-hour training session discussing assessment criteria, benchmarks and possible task expectations.

In the final score for speaking all four tasks are taken into account. Each situational task/item is given a separate score and these together with the scores for other tasks make up the final score for speaking. This study uses the data from the five situational tasks which were the same in all three languages. Performance data of situational tasks in each language was analysed with the Facets package which is based on the Many-Facet Rasch Measurement model (Eckes 2011). This model allows us to consider difference in item difficulty, rater leniency and the functioning of items. Test items are from the NC item bank and have undergone Item Response Theory based analyses which indicate that the tasks function well in all tests.

The data were gathered from three different language tests administered during the autumn of 2011. As the sample for the Swedish test was fairly small in number compared with the English ($N = 215$) and Finnish ($N = 1084$) tests, additional data were gathered during the 2012 spring test bringing the total sample for the Swedish test to 270. The data-collection sample for all languages includes only the test takers who completed the background information sheet.

A background information questionnaire is distributed with every test. It is not compulsory to fill in the information but typically the response rate is around 95%. In addition to the common biographical and demographic data (gender, first language), the questionnaire covers socio-economic factors (education, occupational status, occupational field), purpose of taking the examination, extent of language studies and self-evaluation of target language skills and language use (with family, friends and acquaintances, reading, writing messages, following the media, transactions, work, study; daily, weekly, monthly, not at all). It should be emphasised that the information is self-reported and in the Finnish and Swedish tests it is given in the candidate's second language.

Background and task performance data were analysed per test (language) using descriptive frequency distribution and percent distribution. Cross-tabulation was used to investigate connections between variables. The chi-square test was used to measure the independence of two categorical variables with significance level of 0.05. Adjusted residuals which are based on the comparison of observed and expected frequencies were used to investigate where possible connections might be found. An adjusted residual that is more than 2.0 indicates that the number of cases with the particular variables is significantly larger than would be expected if the null hypothesis were true, with a significance level of 0.05. An adjusted residual that is less than –2.0 indicates that the number of cases is significantly smaller than would be expected if the null hypothesis were true.

[2] Registration is administered by the Finnish National Board of Education.

3. Results and discussion

3.1. Task performance

In the NC speaking tests the situational tasks appear to be the most challenging overall. Luoma (2004: 158) calls situational tasks mini-simulations of reacting in situations. In practice, in situational tasks, what the candidates produce are one- or two-line samples from a dialogue, though in the test situation they are carried out as monologues. This presumes a clear definition of the context from the task, because the test taker has to jump into the 'conversation' for a line or two. The tasks only simulate a dialogue, which raises the question of how interactional these tasks are in terms of authenticity.

Task difficulty varies between test languages and, for instance, the task that was the most difficult for the English test takers was the easiest for the Finnish test takers. Explanations for this variation between the three tests and performance per task within tests can be many. As acknowledged in sociolinguistic-based second language acquisition (SLA) research, second language data do not represent a static phenomenon even at a single point in time. Many external variables, such as the specific task required of a learner, the social status of the interlocutor, gender difference and so forth, affect learner production (Gass, Selinker 2001: 222). This discussion aims to shed light on some issues that emerge from the current data.

Table 1. Distribution of scores in the five situational tasks (%)

	<B1			B1			B2		
	En	Swe	Fi	En	Swe	Fi	En	Swe	Fi
Situation 1	7.9	20.4	40.4	43.7	49.8	39.1	48.4	29.8	20.5
Situation 2	7.0	27.1	24.8	40.0	45.5	54.8	53.0	27.4	20.4
Situation 3	7.4	30.0	43.2	47.0	46.4	38.2	45.6	23.6	18.6
Situation 4	13.0	16.5	30.5	34.0	53.0	47.6	53.0	30.5	21.9
Situation 5	15.3	29.1	18.3	42.3	47.1	61.4	42.3	23.8	20.3

According to test results in the five speaking tasks in the three languages, the English test candidates performed the best (highest level of B2 and lowest of below B1) regardless of the task. The results for the Finnish and Swedish tests are more similar in distribution of grades, i.e. the typical score is B1, but Swedish test candidates performed slightly better overall with more scores of B2. One likely factor behind this result is the language of the written instructions and prompts, which in the Swedish and Finnish tests is the target language but in the English test is the L1 of the test takers. The recommendation is that the tasks should be easy to read. After all, preparation times are short and the tasks are not meant to directly measure the test takers' reading skills. The choice of words can affect the level of difficulty in the task (for instance, 'persuade' in task 1). In general, the instructions and prompts should be simpler than the expected performance of the examinee (Luoma 2004: 169). As the general aim is to have equal interpretation of proficiency levels in all three languages, it is important to notice that understanding the written instructions and prompts may be easier for those taking the English test than for those taking the Swedish and Finnish tests. On the other hand, those taking the English test have to use translation in production while those taking the Swedish and Finnish tests may be able to use at least some of the words directly from the written prompt. The instructions heard from the pre-recorded master tape in the target language only refer to the beginning and ending times of the task.

Situational tasks with brief prompts give a limited definition of the context and test takers have to use their imagination. In test situations, the speakers usually notice such task features that are important and meaningful to them and interpret the tasks in their own way. Thus the products are different and it cannot be expected that all candidates react to a given situation in the same way (Douglas 1998). Situational tasks also require quick reaction and response and may involve adopting a role of some kind. The test taker may or may not be familiar with the role. As the five tasks are different and are presented in quick succession they also require a quick change of roles, which may be both stressful and demanding for candidates.

Knowledge of vocabulary, phrases and idiomatic expressions makes it easier for candidates to succeed in the situation, particularly on level B1. Real life experience in similar speaking situations may not always be sufficient, as the skills of improvising and negotiating meaning are also required (e.g. Bygate 1987: 29). It must be pointed out that all five situational tasks discussed in this study require both routine expressions and improvisation and cannot be responded to by using routine expressions alone. Congratulating in task 4 is well suited for those candidates who know the common expression for the situation, and in task 5 some common adjectives such as *friendly, fair* are sufficient. Also, task 2 may be easy to pass at level B1 with fairly easy language: *I write words many times, I read words* etc. In task 1 test takers talked about their own hobby which can be regarded as a basic level skill but the second part of the task ('tempting') required more intermediate level ability. Situation 3 is not a routine one and seemed to require improvisation more than the others.

It is also important to bear in mind that situational tasks may not be interpreted as authentic and relevant everyday tasks by all candidates. For instance, the tasks may be more unfamiliar to immigrants than they are to those with a Finnish background (e.g. having hobbies). The tasks which simulate a work situation (tasks 3 and 5) may be more difficult for those who do not have much experience of working life in Finland.

3.2. Background factors and test performance

3.2.1. Age

While age is an important factor in the field of language learning and acquisition, for the current study it was found to be of little significance. The NC is a testing system targeted at adults as a way of having their language skills assessed regardless of how the skill has been acquired (through formal study or practical experience). While the age of test takers varies in a similar fashion in each of the test populations (Finnish 18–68, English 20–63, Swedish 15–72) there is some difference in which age group is the most represented. For Swedish and Finnish the largest test taker age group is 31–40 years, to which 31% and 33% (respectively) of our sample belonged. For English, on the other hand, the largest age group is 21–30 years (45%). The younger age of the English test takers relates to the test being used as proof of English for deployment in international military cooperation. A typical English test candidate is a young Finnish male.

Our cross-tabulations show that age appears to be connected to performance in three test tasks but for different language tests. Younger age[3] is associated with better performance in talking about hobbies (task 1) in Finnish ($\chi^2 = 25.7$, $df = 10$, $p = 0.004$) and talking about learning vocabulary (task 2) in Swedish ($\chi^2 = 19.9$, $df = 10$, $p = 0.03$). On the other hand, the older age groups managed better in the congratulating of friends (task 4) ($\chi^2 = 30.0$, $df = 10$, $p = 0.001$). While age was found to correlate with performance in these few instances the connection does not warrant making strong conclusions based on it. In some tasks some age groups performed slightly better than the statistical model expects but this cannot be shown to be at the expense of other age groups, i.e., that a task would be more suited to the younger than the older speaker.

3.2.2. Gender

Literature on issues of language and gender is vast, but there is very little that researchers actually agree on. However, Holmes (1998) has formulated a list of 'sociolinguistic universal tendencies' and suggests that women tend to focus on the affective functions of an interaction, use linguistic devices that stress solidarity, interact in ways that maintain or increase solidarity more often than men, and are stylistically more flexible than men. The situational tasks in this study did not aim at investigating gender difference, but some indication of difference in performance could be seen to link with these tendencies.

In the current data some correlations were found between gender and performance. In the Finnish test women performed better than men in task 1 ($\chi^2 = 21.5$, $df = 2$, $p = 0.0001$), task 2 ($\chi^2 = 25.7$, $df = 2$, $p = 0.0001$) and task 3 ($\chi^2 = 13.6$, $df = 2$, $p = 0.001$). For English women performed better in task 2 ($\chi^2 = 6.5$, $df = 2$, $p = 0.04$) and task 3 ($\chi^2 = 9.4$, $df = 2$, $p = 0.009$). It must be noticed, however, that the English test candidates are a relatively homogenous group (young, male, with relatively high education) with only 14% ($N = 31$) of test takers being female. This can create issues with interpretations of the statistical model, as quantitatively speaking, 31 female test takers cannot be considered representative of the larger population. However, in this data, in tasks 2 and 3 women performed better than men. In the Swedish and Finnish tests gender distribution is more even. In Swedish no gender effect was found.

On one hand we could see the interactional and emotive functions in tasks 1, 2 and 3 to relate to the solidarity and affection tendencies accredited to women but, on the other hand, the tasks also have an informative function. The result then supports the universal tendencies of gender only in part.

[3] E.g. In Finnish, task 1, fewer 21–30 year olds scored below B1 and more of them scored B1 than predicted by the statistical model.

3.2.3. Education

Level of education has been associated with language skills and performance, but again the connection is not simple. Education level is inevitably related to time (age and time spent studying) not to mention the sociocultural, attitudinal and motivational factors that are involved. In this data the cross-tabulation results of education level and situational task performance are inconclusive.

Among the Finnish test candidates a correlation was found between compulsory education (9 years of schooling) as the highest level of education and poorer performance in task 1 (χ^2 = 61.6, *df* = 10, *p* = 0.001), task 2 (χ^2 = 61.2, *df* = 10, *p* = 0.0001) and task 4 (χ^2 = 56.6, *df* = 10, *p* = 0.0001), while university education is associated with scores of B1 in the same tasks and polytechnic[4] education with scores of B2. The education levels of the test takers in the Finnish test cover the range from compulsory education to university education and this connection with low education level and poor performance is in that context logical. On the other hand, among the English test candidates, a slight dependence was found between vocational school education (high school level) and poorer performances (below B1) in most tasks. It should be noted though that this was indicated by the cross-tabulation residuals and not the chi-test which would have been affected by the homogeneity of the population. In this sample only 2% reported having the lowest education level i.e. compulsory education only. In years of study in the Finnish education system vocational school and high school are equal (12 years of schooling), so the difference is not in time of study but something else, which in this data appears in the education level cell. In the Swedish test, no dependence was found between education and performance.

3.2.4. Employment status

The background information sheet asked test takers to indicate their employment status, i.e., whether they are employed, self-employed, unemployed, student, student in labour market training, pensioner, stay at home parent or something else, which they could specify under the category ‘other’. The assumption is that in the Finnish context this would have an effect on target language use opportunities and perhaps performance.

Among candidates for the Finnish test there was a correlation between being employed and performing better in the tasks. There was a clear finding of employed people gaining more scores of B2 in all tasks[5]. This is logical and relates to data discussed in the next section on language use. In the Swedish test, the effect was not as clear, but for tasks 2 (χ^2 = 30.8, *df* = 14, *p* = 0.006) and task 4 (χ^2 = 32.4, *df* = 14, *p* = 0.004) a correlation was found between being employed and better performance was found. There was also a connection between being unemployed and scoring below B1 for task 3 based on adjusted residual results. For the English test no dependence was found, but since the vast majority of candidates for this test are ‘employed’, with only a small percentage reporting any different status, the question does not appear relevant.

[4] In the Finnish system, polytechnic refers to tertiary level education: institutes of technology and universities of applied sciences.

[5] Task 1 χ^2 = 49.3, *df* = 14, *p* = 0.0001, task 2 χ^2 = 28.3, *df* = 14, *p* = 0.01, task 3 χ^2 = 41.8, *df* = 14, *p* = 0.0001, task 4 χ^2 = 55.7, *df* = 14, *p* = 0.0001, task 5 χ^2 = 47.6, *df* = 14, *p* = 0.0001.

3.2.5. Language use

Language use as reported by our sample reflects the Finnish context and the role of the target language as foreign or second language. The language use profile of candidates for the Finnish test indicates that for the majority the target language is used *almost daily* in the public domain: media 70%, reading 58%, transactional activities 66%, work 60% and education 58%. Target language use with family is divided more evenly on the scale, so that 42% report using Finnish almost daily and 30% not at all. Home is the most important domain for first language maintenance in the immigrant context which is also apparent in these language use figures. Finnish use with friends, on the other hand, is frequent; 57% reported almost daily use of Finnish in this domain.

Candidates for the English test present a completely different language use profile. The domain with most frequent use of English is following the media (54% almost daily) while family and transactional activities typically have no use of English (72% and 62% respectively). English use in education is spread more evenly with 30% claiming almost daily use and 34% no use. Language use with friends and acquaintances is the private domain where English is used to some extent, 53% reporting monthly use.

The language use profile of candidates for the Swedish test should be interpreted with caution. Since the population consists of both second language learners and learners of the second national language, the language use profile is an average of two potentially very different profiles. In the private domain, 24% report to using Swedish with the family almost daily and 50% not at all. In the friendship domain, 33% report almost daily use of Swedish and 17% no use. Daily Swedish use to follow the media is not as high as in the other languages: 42%, but on the other hand, only 5% do not follow Swedish media at all. Transactional activities (29% almost daily, 30% no use) and education (33% almost daily, 42% no use) as Swedish use domains are more diverse. Swedish use in the work domain could have been more frequent to reflect the cohorts as expected. After all, the typical motivation for taking the test in Swedish as a second national language is for work purposes and, on the other hand, a Swedish-as-a-second-language learner could be expected to use Swedish also as their work language, if they had chosen Swedish as the official language to test for immigration purposes. Nevertheless, in the current data 44% reported using Swedish at work almost daily and 22% not at all.

Cross-tabulations of language use data and task scores revealed various correlations. Among the Finnish test takers only four language use contexts out of the possible eight[6], showed a correlation with the tasks. Daily use of Finnish with friends linked with higher scores for task 1 ($\chi^2 = 17.0$, $df = 6$, $p = 0.009$), task 4 ($\chi^2 = 15.9$, $df = 6$, $p = 0.02$) and task 5 ($\chi^2 = 19.4$, $df = 6$, $p = 0.004$). Daily use of Finnish in the work domain was connected with better performance in task 3 ($\chi^2 = 22.6$, $df = 6$, $p = 0.001$), task 4 ($\chi^2 = 20.0$, $df = 6$, $p = 0.003$) and task 5 ($\chi^2 = 30.1$, $df = 6$, $p = 0.0001$). There was also dependence between task 5 and daily use of Finnish with the family ($\chi^2 = 16.0$, $df = 6$, $p = 0.01$) and for study ($\chi^2 = 13.7$, $df = 6$, $p = 0.03$). The finding for language use with study is unexpected, since it indicates that test takers who use the target language less for this purpose perform better in the task.

[6] Language use contexts: family, friends and acquaintances, reading, writing messages, following the media, transactions, work, study.

It must be noted though that the way test takers interpreted the category ‘study’ can vary. It is likely that many who are in fact unemployed have stated that they study because they take part in labour-market training. Thus the study they refer to when reporting language use is in fact Finnish language studies.

For the English test data, connections with reported language use are much fewer. It does not appear that the overall strong performance of the English test takers, when compared with results from Finnish and Swedish tests, is influenced by language use. Only daily English use in the contexts of study and work was correlated with performance in the tasks. Good performance in task 2 correlated with daily use of English for study. There was a correlation between weekly English use at work and scores of B1 and not using English at work with scores below B1. Interestingly good performance in the description task (task 5) was only connected with daily English use for study. Very limited or no use of English in many, and particularly the private domains, does not appear to be connected with performance in these tasks. As above with the English data, these dependences are based on adjusted residual results.

For the Swedish test data, connections with candidates’ reported language use are almost nonexistent. We suspect that the combined statistics of two different cohorts flatten the data into being less representative of either separate group.

Overall then language use in the family and friendship domains and, on the other hand in the work and education domains, has a connection with performance in these speaking situation tasks and in these language tests. This makes sense considering that the topics are from these domains. Although target language use, which in the case of second language learners is at a different level to foreign language learners, helps and prepares for the communicative test tasks, it is possible to acquire the relevant skills also through formal learning as demonstrated by the candidates for the English test.

4. Conclusions (and recommendations)

The variation in results of rated performances in the three test languages in the five situational tasks was not great. Analysis of performance data with the Facets package indicates that the test items function adequately. Dependencies were found between rated performance in individual test situations and candidates’ background factors, but these do not appear systematic across languages and do not, in the light of test success data, represent cases of bias. However, some differences warrant further discussion.

In light of the differences between foreign and second language learner performance it was surprising that English test takers performed so well in these situational tasks. Finland has a long tradition of classroom foreign language teaching/learning. Traditionally the formal setting has not supported everyday language use and the learners are not immersed in the target language community. However, in contemporary Finland, English is encountered on a daily basis through audio-visual mass media and various forms of popular culture. Still everyday language use situations may not be familiar unless they are practised as part of formal learning. Learning a second language while immersed in the language community implies also acquisition and ablity to use the language in a natural environment. Considering that for candidates for the Finnish and (some candidates for the) Swedish test the target language is a second language, the situational tasks should have favoured them (e.g. Sajavaara 1999).

Though some correlations were found between background information given by the test takers and situational task performance, they were not so significant as to have been crucial in passing the test. It is more likely that success and failure are in fact influenced by many background variables simultaneously rather than one single piece of information. The test taker's education, cultural background and frequency of language use may be important factors in determining success. For instance, it is no surprise that Estonian candidates did well in the Finnish test as they have the advantage of speaking a language which belongs to the same language family as Finnish. It is typical that Finnish learners with Estonian as their L1 learn very fast at the beginning stages of language learning and easily gain a level of proficiency which is not gained by someone with a more distant L1 background, though with the same duration of learning the language (Jantunen 2011, also Suni 1996). Spaan (2000: 35) has noticed that the results of a language test are influenced by the distance between the target language and the test taker's own language, the familiarity of the test form and test taker's education.

In general, situational tasks are best suited for testing the speaker's pragmatic, sociolinguistic and sociocultural knowledge and skills (register, politeness, social relations, idiomaticity etc.). Often expressions of emotions (regret, gratitude, negative and positive feelings) presume knowledge of certain phrases and the pragmatic features of language. These are also affected by the test takers' own social and cultural context. For instance, in the Finnish test, Swedish and Estonian candidates perform better than other candidates perhaps because their social context is closer to the Finnish one. To minimise these factors in a language test, construct and content validity, reliability and test usefulness should be focused on.

Passing the test is also always influenced by many other factors including the topic, the purpose of communication and the test taker's language learning history, as well as the physical situation, the channel and the test taker's own expectations (see e.g. Huhta 2010: 56). As the majority of Finnish and a part of the Swedish test takers took the test for immigration purposes, the stress in the situation is not without its effect in test behaviour. For instance, being in a language lab may be unfamiliar to some of these candidates. High-stakes tests – important to the test takers' future life – usually represent the value system and practices of the target culture and are often designed for a very homogenous target group. Though the NC is a proficiency test which can be taken by any person with no specified language learning background, and the test is not designed for any type of candidate in particular, the test still carries Finnish language testing conventions and traditions (e.g. use of the language lab), which may in part explain why the candidates for the English test performed so well. This raises the issue of fairness in the language test (Messick 1996, Shohamy 2000, Tarnanen, Mäntylä 2006: 116, 120).

The language of task instructions and prompts is significant to the speaking test. Second language learners are expected to have good reading and also listening comprehension skills in addition to speaking. How important understanding the task rubric is for passing the situation is difficult to ascertain on the basis of test results only, but we may assume that it has some significance particularly for those whose comprehension skills are not very good.

The idea of authenticity is regarded as important for the construct validity of any language test. The aim is to write as authentic situational tasks as possible, but one must bear in mind that external similarity between a test task and a communicative situation does not necessarily fully reflect the nature of the situation. To be authentic, test tasks should be meaningful to the test takers, challenging enough, and engage them in activities that mirror their real-life contexts (e.g. O'Malley, Valdez Pierce 1996: 5, Douglas 2000: 18). The important issue here is perceived relevance. If we wish to have the test takers perform at their very best in the test situation, we should relate the test tasks in type and topical content to the target language use tasks outside the test situation. To be able to make interpretations on the test task and its correspondence with the 'real world' we also need information on how the test taker processes language and how he/she approaches the task (Bachman 1990, Huhta, Takala 1999).

Since the language test is aimed at any adult regardless of their language learning background, the test results indicate that the situational tasks discussed in this study are well suited for measuring test takers' ability to use spoken language in various language use situations.

References

Alderson, Charles J.; Clapham, Caroline; Wall, Dianne 1995. Language Test Construction and Validation. Cambridge: Cambridge University Press.

ALTE Manual 2011 = Manual for Language Test Development and Examining. For use with the CEFR. Produced by ALTE [Association of Language Testers in Europe] on behalf of the Language Policy Division, Council of Europe, April 2011.

Bachman, Lyle F. 1990. Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Bachman, Lyle F.; Palmer, Adrian S. 1996. Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: Oxford University Press.

Bygate, Martin 1987. Speaking. Oxford: Oxford University Press.

Canale, Michael; Swain, Merrill 1980. Theoretical Bases of Communicative Approches to Second Language Teaching and Testing. Applied Linguistics 1 (1). Oxford: Oxford University Press.

CEFR 2001 = Common European Framework of Reference: Learning, Teaching, Assessment 2001. Council of Europe. Cambridge: Cambridge University Press.

Douglas, Dan 1998. Testing methods in contex-based second language research. – Lyle F. Bachman, Andrew D. Cohen (Eds.). Interfaces Between Second Language Acquistion and Language Testing Research. Cambridge: Cambridge University Press, 141–155.

Douglas, Dan 2000. Assessing Languages for Specific Purposes. Cambridge: Cambridge University Press.

Eckes, Thomas 2011. Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments. Frankfurt am Main: Peter Lang.

Gass, Susan M.; Selinker, Larry 2001. Second Language Acquisition: An Introductory Course. 2nd edition. Mahwah, NJ: Lawrence & Erlbaum Associates Inc.

Holmes, Janet 1998. Women's talk: The question of sociolinguistic universals. – Jennifer Coates (Ed.). Language and Gender: A Reader. Oxford/Malden, MA: Blackwell, 461–483.

Huhta, Ari 2010. Suullisen kielitaidon arviointi: Mitä, miten ja miksi – ja voiko Eurooppalainen viitekehys auttaa siinä? – Sabine Grasz, Joachim Schlabach, Edeltraud Sormunen, Ari Huhta (Hrsg.). QualiDaF – Qualitätssicherung, Lernziele und Beruteilungskriterien für den fachbezogenen Deutschunterricht. Jyväskylä: Jyväskylän yliopisto, Soveltavan kielentutkimuksen keskus, 31–56.

Huhta, Ari; Takala, Sauli 1999. Kielitaidon arviointi. – Kari Sajavaara, Arja Piirainen-Marsh (Toim.). Kielenoppimisen kysymyksiä. Soveltavan kielentutkimuksen teoriaa ja käytäntöä 7. Jyväskylä: Jyväskylän yliopisto, Soveltavan kielentutkimuksen keskus, 179–228.

Jantunen, Jarmo H. 2011. Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. – Lähivõrdlusi. Lähivertailuja, 21, 86–105. http://dx.doi.org/10.5128/LV21.04

Luoma, Sari 2004. Assessing Speaking. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511733017

Messick, Samuel 1996. Validity and washback in language testing. – Language Testing, 13 (3), 241–256. http://dx.doi.org/10.1177/026553229601300302

O'Malley, Michael J.; Valdez Pierce, Lorraine 1996. Authentic Assessment For English Language Learners. Practical Approaches For Teachers. Addison-Wesley Longman.

Sajavaara, Kari 1999. Toisen kielen oppiminen. – Kari Sajavaara, Arja Piirainen-Marsh (Toim.). Kielenoppimisen kysymyksiä. Soveltavan kielentutkimuksen teoriaa ja käytäntöä 7. Jyväskylä: Jyväskylän yliopisto, Soveltavan kielentutkimuksen keskus, 73–102.

Shohamy, Elana 2000. Fairness in language testing. – Antony J. Kunnan (Ed.). Fairness and Validation on Language Assessment. Selected papers from the 19th Language Testing Research Colloquim, Orlando, Florida. Studies in Language Testing 9. Cambridge: Cambridge: Cambridge University Press, 15–19.

Spaan, Mary 2000. Enhancing fairness through a social contract. – Antony J. Kunnan (Ed.). Fairness and Validation on Language Assessment: Selected papers from the 19th Language Testing Research Colloquim, Orlando, Florida. Studies in Language Testing 9. Cambridge: Cambridge: Cambridge University Press, 35–38.

Suni, Minna 1996. Maahanmuuttajaoppilaiden suomen kielen taito peruskoulun päättövaiheessa. – Opetushallituksen moniste 11/1996.

Tarnanen, Mirja; Mäntylä, Katja 2006. Toisen ja vieraan kielen oppijat yleisissä kielitutkinnoissa. – Päivi Pietilä, Pekka Lintunen, Heini-Marja Järvinen (Toim.). Kielenoppija tänään – Language Learners of Today. AFinLa vuosikirja 2006. Suomen soveltavan kielitieteen yhdistyksen julkaisuja 64. Jyväskylä: Jyväskylän yliopistopaino, 105–123.

The Finnish National Board of Education and the University of Jyväskylä 2002. The Framework of the Finnish National Cerficates. Helsinki: Edita Prima Oy.

The Finnish National Board of Education and the University of Jyväskylä 2011. The Framework of the Finnish National Cerficates. Tampere: Tampereen yliopistopaino.

Weir, Cyrill J. 2005. Language Testing and Validation: An Evidence-Based Approach. Basingstoke: Palgrave Macmillan.

Sari Ahola (Centre for Applied Language Studies, University of Jyväskylä, Finland) is involved in language assessment and activities for the development of the Finnish National Certificates of Language Proficiency.
sari.ahola@jyu.fi

Tiina Lammervo (Centre for Applied Language Studies, University of Jyväskylä, Finland) is involved in language assessment and activities for the development of the Finnish National Certificates of Language Proficiency.
tiina.lammervo@jyu.fi

Reeta Neittaanmäki (Centre for Applied Language Studies, University of Jyväskylä, Finland) is involved in language assessment and activities for the development of the Finnish National Certificates of Language Proficiency.
reeta.neittaanmaki@jyu.fi

Sari Ohranen (Centre for Applied Language Studies, University of Jyväskylä, Finland) is involved in language assessment and activities for the development of the Finnish National Certificates of Language Proficiency.
sari.ohranen@jyu.fi

Henna Tossavainen (Centre for Applied Language Studies, University of Jyväskylä, Finland) is involved in language assessment and activities for the development of the Finnish National Certificates of Language Proficiency.
henna.tossavainen@jyu.fi

KOLME ERINEVA KEELETESTI RÄÄKIMISOSADE VÕRDLUSTULEMUSED

Sari Ahola, Tiina Lammervo, Reeta Neittaanmäki, Sari Ohranen, Henna Tossavainen
Jyväskylä Ülikool

Artikkel tutvustab uuringut, milles võrreldi õppijate tulemusi kolme erineva keeleeksami rääkimise allosas. Tulemusi püütakse seletada eksaminandide poolt antud info kaudu nende demograafilise, haridusliku, professionaalse ning keeleõppe tausta kohta. Uuringu eesmärk oli panustada eksamiülesannete konstrukti (mõõdetava omaduse) valiidsusse ja selgitada välja, kas ülesanded olid ühtmoodi valiidsed kõigi kolme keele puhul.

Uuringus viidi 2011. aasta sügisel läbi katse, mille käigus võrreldi rääkimise osa Soome riiklikel kesktaseme eksamitel (*The National Certificate of Language Proficiency*): soome, rootsi ja inglise keeles. Eksaminandidele anti selleks viis sarnast suhtlussituatsiooni. Rootsi keele testi puhul koguti andmeid veel ka 2012. aasta kevadel. Tulemusi ja taustafaktoreid analüüsiti sagedus- ja protsentjaotusandmete põhjal, muutujatevahelisi seoseid risttabelite abil.

Üksikvastuste teooria (*Item Response Theory*) põhine analüüs näitas, et eksamikorraldaja küsimustepangast valitud ülesanded toimivad kõigi kolme keele puhul hästi nii küsimuste raskuse kui ka eristusvõime osas. Uuringu käigus selgus aga, et selleks, et piisava adekvaatsusega kirjeldada eksaminandide tulemuste seost nende taustainfoga, on vaja enam andmeid.

Uuringu tulemused näitavad, et eksami sooritamise edukus sõltub tõenäoliselt paljudest taustamuutujatest ning muudest testi situatsiooniga seotud muutujatest. Ilmnes ka sõltuvusi eri testsituatsioonides saadud tulemuste ja kandidaatide taustatunnuste vahel, kuid need ei esinenud eri keelte testide vahel süstemaatiliselt ning ei ole seetõttu selgelteristuvaks eduka testisoorituse mõjufaktoriks. Siiski võivad eksaminandi haridus, kultuuritaust ja keelekasutuse sagedus olla olulisteks soorituse edukust mõjutavateks teguriteks.

Võtmesõnad: teise keele ja võõrkeele testimine, suuline keel, valiidsus, soome keele eksam, rootsi keele eksam, inglise keele eksam