doi:10.5128/ERYa10.17

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 10, 269–281

INVESTIGATING CULTURAL VARIABILITY IN RATER JUDGEMENTS OF ORAL PROFICIENCY INTERVIEWS

Irina Stassenko, Liljana Skopinskaja, Suliko Liiv

Abstract. The article is concerned if the cultural background of non-native raters could lead to substantial differences in the assessment of students' oral performances during the national examination in the English language in Estonia. The discussion involves the analysis of the ratings of twenty video-recordings of oral proficiency interviews by two rater groups of the Estonian and Russian origin, and a questionnaire study investigating rater perceptions of the national examination. Despite the lack of any marked cultural bias being displayed in the raters' behaviour, the results of the study reveal a number of significant differences in their perceptions of various aspects of the national examination as well as their own behaviour during the rating process.*

Keywords: assessment, rater variation, rater reliability, cultural validity, oral proficiency interview

1. Introduction

The assessment of oral proficiency interviews can rarely be error-free as scores assigned to test takers are based on a reflection, not only of the quality of a candidate's performance, but also of the qualities of the rater (McNamara 2000: 37). Extensive research into rater behaviour indicates different factors affecting the variability in raters' judgements, such as severity, i.e., how severe or lenient a rater is (Lumley, McNamara 1995: 56, Bachman et al. 1995: 238–257), consistency in assessment (Wigglesworth 1993: 308–309, Luoma 2004), variety of perceptions of what constitutes speaking proficiency (Pollitt, Murray 1996, Fulcher 2003: 144), educational level, i.e., whether the rater is a teacher or a professional rater (Brown 1995, Fulcher 2003: 142, Pajupuu 2007) and linguistic background, i.e., whether the rater is a native or a non-native speaker (Brown 1995, Winke et al. 2012).

* This research has partly been supported by the Estonian Science Foundation Grant No 9037 "Assessing Variables: Creating Benchmarks in High-Stakes Proficiency Tests".

While most of the studies on raters' behaviour concentrate either on their linguistic background or professional experience, little attention has been paid to non-native rater variation, although some cultural variability has been detected in interlocutors' handling of oral proficiency interviews in E. Alas' doctoral thesis (2010). In addition, cultural variation may be observed in the diversity of communication styles, talk distribution, or turn-taking patterns (Tannen 1984, Trompenaars 1994). In other words, communication styles reflect cultural values and the different ways in which interpersonal relations are believed to be achieved best (FitzGerald 2003: 79). Hence, the concept of cultural validity, as a form of test validity, has recently been introduced to link cultural and linguistic factors to test takers' assessment outcomes (Abedi 2011). Based on the differences in value dimensions as well as communication styles between the two cultures, where Estonian culture is characterised as an individualist, small power distance and low uncertainty avoidance culture, as contrasted to Russian culture with its collectivist, large power distance and high uncertainty avoidance orientations (Hofstede 1997, Pajupuu 2001), it is possible to assume that the raters' ways of thinking that permeate their culture and predisposed notions of what good language proficiency constitutes may also influence their rating behaviour, thus calling into question the reliability of test results. Moreover, serious doubts have lately been raised by INNOVE, the former National Examinations and Qualification Centre of Estonia, as to rater validity and reliability in oral proficiency interview results of Russian-based school learners, as compared to their much lower scores in the other parts of the national examination in English (Kriisa 2012: 27).

In view of the absence of relevant studies conducted in Estonia the following research hypothesis was formulated: the conduct of the two non-native groups of Estonian and Russian raters in the oral part of the national examination in English in Estonia will display culture-related differences in their rating process that may in turn affect the candidates' scores.

2. Method and participants

As a rule, raters involved in the speaking test of the national English language examination in Estonia are practising teachers who have undergone standardised assessor training. The speaking test to be marked comprises an introduction (not rated), a monologue and follow-up questions (task 1) and a role play (task 2). The marking scale for speaking consists of four categories (communication, vocabulary, grammar, and fluency & pronunciation) in terms of which the language produced by a candidate is described (REKK 2008). The marking scale has six levels of language proficiency with the descriptors of what the candidate is supposed to be able to express under each of the abovementioned four categories. The minimum score for each of the criteria is 0, and the maximum score is 5 points. The maximum total score that can be assigned is 20 points. The candidate's final mark represents a sum total of all the aforementioned categories. However, for the sake of the current study the raters were required to evaluate both speaking tasks – a monologue and a role play – separately, although only one grade is usually assigned for the speaking test.

18 raters were randomly selected from Estonian and Russian-based schools to mark 20 video-recordings of oral proficiency interviews produced during the national English language examination of 2012 within the framework of the Estonian Science Foundation Grant No 9037. Most of the teachers participating in the study had an assessment experience in the national examination, the length of service ranging from 2 to 18 years. Out of 18 assessors, only 4 (2 Estonians and 2 Russians) were short of relevant experience. As to their teaching experience, the length of service of the Estonian teachers ranged from less than 5 years to more than 15 years, while the Russian teachers had mostly more than 15 years of service.

Although 7105 students participated in the national English language examination in the spring of 2012 (Kriisa 2012: 3), the choice of the material for the study was limited as the video-recordings were made only with the written consent of all the parties: a candidate, an interviewer and a rater. Thus the permission for recording was obtained only from 20 Estonian 12th-formers studying in one Tallinn and two Harju county secondary schools. In spite of the absence of Russian test-takers, which is a drawback of the present study, the recorded performances represented all levels of oral proficiency in English with the students' scores ranging from the highest to the lowest ones. The database of 20 video-recordings comprises 300 minutes of interview time in total. The recordings as well as the raters were assigned letter-codes and numbers in a random order: Clip1 to Clip20 for the recordings, E1 to E9 for the raters from Estonian-based schools and R1 to R9 for the Russian-based ones. All the participants were female and non-native speakers of English.

In order to elicit more data about the raters' attitudes to and their behaviour during the assessment procedure, a questionnaire study was implemented among the same participants. The data obtained from the Estonian and Russian raters' assessment sheets and questionnaire forms were analysed with the aim of establishing any distinctive differences in the assessment of the aforementioned cultural groups. This was further analysed by statistical methods (ANOVA, the 95% Confidence Interval for Mean, Chi-Square Tests) to ascertain if the collected variations were statistically relevant.

3. Discussion and results

The section presents the findings of the two research instruments obtained from the assessment of the oral performance of 20 test-takers by the two rater groups, and a questionnaire study conducted among the same assessors.

3.1. Assessment of oral proficiency interviews

The results of the assessment of each task of the speaking test are analysed by means of juxtaposition of the ratings of the two cultural groups to discover any differences in the respective assessment procedures.

3.1.1. Monologue and follow-up questions (task 1)

The analysis of the marks assigned for the first task is based upon the four criteria of the marking scale (REKK 2008). Table 1 below reflects the mean scores assigned for each aspect of the monologue, these slightly exceeding 4 points.

Table 1. The mean scores for task 1

Nationality	Communication	Vocabulary	Grammar	Fluency & pronunciation
Estonian	4.36	4.14	4.04	4.36
Russian	4.41	4.31	4.20	4.39

Although no statistically significant differences were revealed in the distribution of the scores afforded for the candidates' communication skills and fluency & pronunciation, since more than 55% of the 20 performances received the highest scores for these aspects, there is slight variation in the raters' opinions as regards the candidates' grammar and vocabulary usage, this demonstrating the respective raters' different perceptions of what constitutes good knowledge of grammar or vocabulary. The Estonian raters were more reluctant to award either the lowest or the highest scores for these aspects. Only 29% of the test-takers received the highest score for their grammar by the Estonian rater group, whereas with the Russian group the number of such candidates rose to 41%. The scores for vocabulary usage were 32% and 47% respectively.

More controversy emerged in connection to three recordings (Clip2, Clip9, and Clip11) where the highest scores assigned by the Russian rater group were in marked contrast to those of the Estonian group, the latter evaluating the same students as hesitant speakers with a limited choice of vocabulary and grammar (mark 3).

Despite having different perceptions of the proficiency level of some of the candidates (these being statistically irrelevant), both rater groups emphasized in their notes the importance of a logical structure and varied vocabulary choice in a monologue.

3.1.2. Role play (task 2)

Similarly to the rating of monologues, the mean scores for the second task do not reveal any statistically significant differences, these being even between the two groups while assessing communication skills and fluency & pronunciation, and slightly higher for lexis and grammar usage with the Russian raters (see table 2).

Table 2. The mean scores for task 2

Nationality	Communication	Vocabulary	Grammar	Fluency & pronunciation
Estonian	4.35	4.21	4.00	4.43
Russian	4.38	4.38	4.11	4.45

The same divergence of opinions exists in the assessment of grammar and lexis here: a smaller number of candidates received the highest scores from the Estonian

raters. In comparison to the Russian assessors who awarded 56% of the students with the highest score for their vocabulary, the Estonian raters considered 41% worthy of this mark. With grammar assessment, the percentages were 36% and 26% respectively. Yet the frequency of awarding the lowest score for the candidates' role playing was higher with the Russian raters, as up to 8 % of the candidates received only 1 point for any of the criteria in the marking scale, whereas with the Estonian assessors it was only 3%.

To sum it up, both rater groups share similar perceptions about the candidates' good role playing performance in their notes, emphasizing the presence of both the introduction and conclusion to the role play, natural turn-taking patterns and question formation skills.

3.1.3. The total scores for oral proficiency interviews

After the points for both tasks had been awarded, the raters were to decide upon the final scores. As can be seen in table 3, there are no obvious differences in the mean scores of the two rater groups. Although no apparent severity or leniency is exposed in their behaviour, the results still show that the Russian assessors tended to give slightly higher points for each category, hence their total mean score (17.27) is slightly higher than that of the Estonian raters (16.96). The 95% Confidence Interval for Mean does not, however, show any notable discrepancies between the groups.

Table 3. The overall mean scores for tasks 1 and 2

Nationality	Communication	Vocabulary	Grammar	Fluency & pronunciation
Estonian	4.37	4.18	4.02	4.39
Russian	4.46	4.32	4.14	4.43

As to the frequency of awarding minimum and maximum points, the Russian raters were more critical of the candidates' communication skills and vocabulary usage by giving the lowest mark for these, whereas the minimum score assigned by the Estonian raters for the same categories was 2 points. In case of fluency & pronunciation, the score distribution was the reverse: 1 point received from the Estonian raters and 2 points from the Russian ones (see table 4).

Table 4. Maximum and minimum points awarded for each aspect of the candidate's performance

Task	Nationality	Minimum	Maximum
Communication	Russian	1	5
	Estonian	2	5
Vocabulary	Russian	1	5
	Estonian	2	5
Grammar	Russian	1	5
	Estonian	1	5
Fluency & pronunciation	Russian	2	5
	Estonian	1	5

The score distribution between the rater groups was also analysed with respect to their work experience as a national examination assessor and the length of service (see table 5).

Table 5. Correlation between the length of service and the number of scores awarded for fluency & pronunciation (N – the total number of the raters being multiplied by the number of the videorecordings assessed)

Length of service	N	Mean	Std. deviation	Std. error	95% confidence interval for mean		Min	Max
					Lower bound	Upper bound		
1–5	80	4.23	.763	.085	4.06	4.39	2	5
6–10	40	4.55	.639	.101	4.35	4.75	3	5
11–15	40	4.55	.677	.107	4.33	4.77	2	5
|>15	200	4.44	.741	.052	4.33	4.54	1	5

The analysis of the variance does not reveal any significant differences between the awarded scores and the rating experience. However, there is statistically relevant correlation (according to the 95% Confidence Interval for Mean) between the rater's teaching experience and the score awarding for fluency & pronunciation. The mean score (4.23) assigned by less experienced raters with less than five years of service was the lowest in comparison to that of the more experienced teachers. The scores assigned by the most experienced raters with more than 15 years of service was higher (4.44) than the previous group's but lower than that of the participants with 6-10 and 11-15 years of teaching (4.55). Less experienced raters (with 6-10 years of service) tended to give higher scores (3 points) for what the most experienced teachers (with more than 15 years of experience) afforded only the minimum amount of points (1 point), which resulted thus in the confusion, on the part of the former group, while distinguishing between a hesitant speaker and a very laconic speaker, (see REKK 2008) and unfair grading of such learners.

3.2. Questionnaire study results

The aim of the questionnaire study was to elicit data about the assessors' opinions of the quality of the existing marking scale (i.e., whether it needed any further improvement), of the assessment of various aspects of a candidate's oral performance (i.e., the effect of a candidate's accent on the rating, the importance of content, pronunciation, lexis, fluency and grammar use in a candidate's output) as well as of their own rating behaviour (i.e., the existence of any distracting factors affecting the rating, interference of a rater's cultural identity with the assessment, effect of a test taker's self-confidence, willingness for cooperation with the interlocutor, display of interest in the topic, etc.). In addition, the participants were required to provide background information about themselves, as to the length of service as a teacher and national examination assessor, and whether they experienced any need for further improvement of their rating skills.

3.2.1. The raters' impressions of the marking scale

When asked about the raters' perceptions of the existing marking scale for speaking, a notable variance between the two rater groups is exposed in their need for further improvements.There are statistically relevant differences between the groups: 89% of the Estonian raters and only 11% of the Russian ones were in favour of further improvements (see table 6).

Table 6. The necessity to improve the rating scale for speaking

Nationality	Rating scale needs improvement		Total
	0	1	
Russian	88.9%	11.1%	100.0%
Estonian	11.1%	88.9%	100.0%
Total	50.0%	50.0%	100.0%

Unlike their Russian colleagues, the majority of the Estonian raters criticized the existing marking scale for lack of proper criteria in assessing speaking and of appropriate levels for each category. The most controversial aspects in the rating scale referred to the ambiguity in the terms like occasional mistakes, an independent speaker versus a good speaker, and communication versus fluency (see REKK 2008). In other words, the raters felt that some assessment criteria seemed to overlap. The participants also complained of the difficulties while applying the same criteria to assess both the monologue and role playing.

Finally, both groups were also highly critical of the existing marking scale – 67% of all the participants doubted its validity.

3.2.2. Factors affecting the raters' behaviour

The research reveals that even after having passed standardised assessor training 78% of the interviewed raters admitted the presence of a number of factors disturbing the assessment process, in terms of either tasks or procedural matters.

The first objection raised concerns an obligation to record oral proficiency interviews. According to rater R6, this distracts not only a candidate but also an assessor/interviewer who has to worry about the quality of the recording.

In connection with the tasks, the raters opted for more variability in the question and answer section, as sometimes student cue cards contain prompts that are difficult to match with the answers in the interviewer script, this resulting in awkward interaction between the candidate and the interviewer, which is difficult to assess.

The participants were further asked whether their cultural identity may interfere with the assessment process. 94% of all the respondents denied this.

Based on the research by Winke et al. (2012), the raters were also inquired of a possible effect of a candidate's accent on the rating. Strangely enough, neither the Russian (89%) nor the Estonian (67%) rater group had ever experienced such an influence, which may be accounted for by the respondents' insufficient experience with various learner accents. However, both rater groups claimed to lower their

marks in case of serious miscomprehension occasioned by a test taker's strong accent.

Next, it was hypothesized in the study that both rater groups may be affected by factors outside the marking scale, such as their preconceived impressions of what constitutes good speaking proficiency.

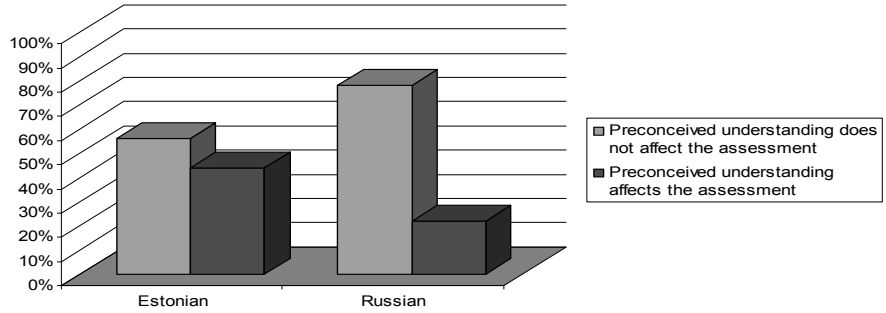

Figure 1. The influence of preconceived understanding of language proficiency levels on the candidates' assessment (x-axis: nationality, y-axis: percentage of raters)

Figure 1 demonstrates slight balance in the opinions of the Estonian rater group where 56% denied and 44% admitted of their rating behaviour being affected by some preconceived understanding of different speaking proficiency levels. The majority of the Russian group denied (78%) such an influence, as only 22% admitted its effect. Some raters were concerned in their notes about their tendency to compare candidates against each other, rather than checking a candidate's skills against the marking scale criteria.

3.2.3. Features of the candidate's language and the rater's behaviour

This part of the questionnaire contained closed-ended questions as to which features of the test takers' ability (content, pronunciation, lexis, fluency, grammar) may affect the scoring process.

Both rater groups (66% of the Estonians and 44% of the Russians) regarded the content to be the most essential aspect in assessing students' oral performances. Next, a wide range of vocabulary was appreciated by 56% of the Estonian raters and 33% of the Russian ones. Strangely enough, the presence of a candidate's proper pronunciation skills was ignored by both group. As to the importance of fluency and grammar, the variance analysis ANOVA demonstrates statistically significant differences between the two groups (see table 7). The mean degrees of the importance of fluency and grammar are 4.000 and 2.56 respectively within the Russian group, thus grammar being emphasized over fluency by them. This means that if a candidate is not able to maintain the flow of speech but nevertheless uses correct grammar structures, this will have no effect on the rater's judgements. In the Estonian group, the figures are 2.833 for fluency and 3.78 for grammar, with fluency being stressed over grammar (see table 7). Furthemore, none of the respondents within this group considered fluency to be the least essential aspect in assessing the speaking skill, whereas 33% thought of grammar as the least important in a candidate's output. Thus the rating of fluency and grammar knowledge in oral proficiency tests is affected by the assessors' cultural background.

Table 7. Comparison of the relevance of fluency and grammar usage in speaking between the two rater groups

Feature	Nationality	N	Mean	Std. deviation	Std. error	95% confidence interval for mean		Min	Max
						Lower bound	Upper bound		
Importance of fluency	Russian	9	4.000	1.000	.333	3.231	4.769	2	5
	Estonian	9	2.833	1.2748	.4249	1.853	3.813	1	4
Importance of grammar use	Russian	9	2.560	1.014	.338	1.780	3.330	1	4
	Estonian	9	3.780	1.302	.434	2.780	4.780	1	5

The raters were further inquired about the most distracting factors in a candidate's speech, such as mispronunciation, hesitation, self-repetition, limited vocabulary choice and grammatical inaccuracy. Contrary to the similarities in the attitude of both groups towards mispronunciation, hesitation and repetition, statistically relevant differences between the groups emerge only with respect to a candidate's vocabulary choice. In contrast to the Russian raters (22%), the Estonian ones (66%) are highly distracted by a candidate's limited wordstock (see table 8). Obviously those candidates whose vocabulary usage is restricted will not obtain high scores from the latter group, notwithstanding a high level of proficiency in the other aspects of their speech (e.g., pronunciation, grammatical accuracy).

Although the data about the last category of the distracting factors, i.e., ungrammatical forms, show no statistically relevant differences between the two cultural groups, the Russian participants admitted this distraction in 56% of all cases. Conversely, the Estonian raters with the same response were represented by 78%. Such a high percentage of the raters admitting the distraction of grammatical inaccuracy seems inconsistent since most of the raters had claimed earlier of the prevalence of content over grammar in assessing the speaking skills.

Table 8. The most distracting factor when rating speaking – vocabulary use

Nationality	Distracting vocabulary use		Total
	0	1	
Russian	77.8%	22.2%	100.0%
Estonian	33.3%	66.7%	100.0%
Total	55.6%	44.4%	100.0%

Finally, the study aimed at analysing whether the raters' impressions may be affected by the following aspects in a candidate's performance, such as the candidate's self-confidence, display of interest in the topic, willingness to interact with the interlocutor, elaboration on one's ideas, and personal response. There were no statistically relevant differences between the groups, apart from the candidate's willingness to interact with the interlocutor. Contrary to the Russian raters (56%), all the Estonians (100%) regarded the test taker's lack of communication strategies as detrimental to his or her positive assessment (see table 9). This may refer to Russian assessors' willingness to overrate test takers in spite of a lower level of

their communication skills. The statistical analysis confirms the significance of this discrepancy between the two cultural groups.

Table 9. The effect of willingness for interaction on the perception of communication and fluency

Nationality	The effect of willingness for interaction on the perception of communication and fluency		Total
	0	1	
Russian	44.4%	55.6%	100.0%
Estonian	0%	100.0%	100.0%
Total	22.2%	77.8%	100.0%

4. Conclusion

The present research is a first attempt to compare the behaviour of the two cultural groups of non-native raters within the context of rating oral proficiency interviews of the national examination in English. The analysis of the assessment of 20 video recordings by 18 Estonian and Russian raters as well as the questionnaire study confirmed the hypothesis to some extent.

In spite of the absence of statistically significant differences in the scores awarded by both groups, the Russian raters tended to assign the highest as well as the lowest points for the four criteria (communication, vocabulary, grammar, and fluency & pronunciation) of the candidates' performance more frequently than their Estonian colleagues. The research revealed significant correlation between the raters' teaching experience and their severity of rating: the assessors with less than 5 years of experience were the strictest in assessing a candidate's fluency & pronunciation while the raters with 6-10 and 11-15 years of service were the most lenient, awarding the highest points for the same criteria.

The data obtained from the questionnaire study refers to the following statistically relevant differences in the perceptions and behaviour of the two rater groups: a) attitude towards the existing rating scale, since the Estonian group was in favour of some improvements to make the rating process more reliable, whereas the Russian one exposed no need for that; b) perception of the degree of relevance of fluency or grammar in the assessment, as the Russians emphasised grammar knowledge over fluency, and the Estonians attached more importance to fluency rather than to grammatical accuracy; c) perception of the most distracting factors in rating oral proficiency interviews, since most of the Estonians (66%) were disturbed by the candidates' limited vocabulary choice, whereas only 22% of the Russians found this distracting; d) attitude towards the candidate's interaction skills where all the Estonians (100%) admitted that their perception of a test taker's fluency and communication skills is affected by his or her ease of interaction with the interlocutor, and the proportion of the Russian raters sharing the same view comprised only 56%.

In view of a number of statistically insignificant differences in the raters' behaviour that emerged in the course of the research, it is difficult to claim whether the elicited differences between the two rater groups may be attributed to their cultural background, or whether they are the result of a limited number of the participants.

Nevertheless, the findings of the study provide valuable insight into rater behaviour which is imperative for developing valid testing tools, training raters and generating sound test results.

References

Abedi, Jamal 2011. Assessing English language learners: Critical issues. – Maria del Rosario Basterra, Elise Trumbull, Guillermo Solano-Flores (Eds.). Cultural Validity in Assessment: Addressing Linguistic and Cultural Diversity. New York: Routledge, 49–71.

Alas, Ene 2010. The English Language National Examination Validity Defined by Its Oral Proficiency Interview Interlocutor Behaviour. Tallinn University Dissertations on Humanities 22. Tallinn: Tallinn University Press.

Bachman, Lyle F.; Lynch, Brian K.; Mason, Maureen 1995. Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. – Language Testing, 12 (2), 238–257. http://dx.doi.org/10.1177/026553229501200206

Brown, Anne 1995. The effect of rater variables in the development of an occupation-specific language performance test. – Language Testing, 12 (1), 1–15. http://dx.doi.org/10.1177/026553229501200101

FitzGerald, Helen Gay 2003. How Different Are We? Spoken Discourse in Intercultural Communication: The Significance of the Situational Context. Languages for Intercultural Communication and Education 4. Clevedon: Multilingual Matters.

Fulcher, Glenn 2003. Testing Second Language Speaking. London: Longman/Pearson Education.

Hofstede, Geert 1997. Culture and Organisations: Software of the Mind. New York: McGraw-Hill.

Kriisa, Kristel 2012. 2012. aasta inglise keele riigieksami lühianalüüs. [Analysis of the National Examination in the English Language of 2012.] http://ekk.edu.ee/valdkonnad/uldharidusvalishindamine/riigieksamite-materjalid-2012/inglise-keel (10.3.2013).

Lumley, Tom; McNamara, Tim F. 1995. Rater characteristics and rater bias: Implications for training. – Language Testing, 12 (1), 54–71. http://dx.doi.org/10.1177/026553229501200104

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

McNamara, Tim F. 2000. Language Testing. Oxford: Oxford University Press.

Pajupuu, Hille 2001. Kuidas kohaneda võõras kultuuris: käsiraamat. [How to Adapt Oneself to a Foreign Culture: A Handbook.] Tallinn: Eesti Keele Sihtasutus.

Pajupuu, Hille 2007. Kuidas hinnata suure panusega testide hindajaid. [How to assess the raters of high-stakes tests.] – Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics, 3, 221–233. http://dx.doi.org/10.5128/ERYa3.15

Pollitt, Alastair; Murray, Neil L. 1996. What raters really pay attention to. – Performance Testing, Cognition and Assessment. Selected Papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem. Cambridge: Cambridge University Press, 74–91.

REKK 2008 = National Examinations and Qualification Centre of Estonia 2008. Marking scale for speaking. www.ekk.edu.ee/valdkonnad/uldharidusvalishindamine/riigieksamite-materjalid-2008/inglise-keel (26.12.2012).

Tannen, Deborah 1984. Conversational Style: Analysing Talk Among Friends. Norwood, NJ: Albex.

Trompenaars, Fons 1994. Riding the Waves of Culture: Understanding Cultural Diversity in Business. London: Nicholas Brealey.

Wigglesworth, Gillian 1993. Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. – Language Testing, 10 (3), 305–319. http://dx.doi.org/10.1177/026553229301000306

Winke, Paula; Gass, Susan; Myford, Carol 2012. Raters' L2 background as a potential source of bias in rating oral performance. – Language Testing, 30 (2), 231–252. http://dx.doi.org/10.1177/0265532212456968

Irina Stassenko (Tallinn Polytechnic School) is a graduate of Tallinn University (2013) who has also been trained as a national examination interviewer.
Pärnu Rd 57, 10135 Tallinn, Estonia
irina.stsnk@gmail.com

Liljana Skopinskaja (Tallinn University), her research interests include intercultural communication, foreign language teaching methodology and language teacher education.
Narva Rd 25, 10120 Tallinn, Estonia
liljana.skopinskaja@tlu.ee

Suliko Liiv (Tallinn University), her research interests are contrastive linguistics, language policy, intercultural communication and teacher education.
Narva Rd 25, 10120 Tallinn, Estonia
liiv@tlu.ee

KULTUURILISTE ERINEVUSTE UURING SUULISE KEELEPÄDEVUSTESTI HINDAJA KÄITUMISES

Irina Stassenko, Liljana Skopinskaja, Suliko Liiv
Tallinna Ülikool

Artiklis analüüsitakse kultuurilisi erinevusi keelepädevustesti hindaja käitumises inglise keele riigieksami läbiviimisel. Autoreid huvitab, kas eesti- ja venekeelsete hindajate käitumises esineb statistiliselt olulisi erinevusi, mis võiksid riigieksami tulemuste usaldusväärsust olulisel määral mõjutada. Uurimuse aluseks on kahekümne inglise keele riigieksami videosalvestuse hindamine, mille viisid läbi kaheksa eesti- ja kaheksa venekeelset hindajat, ja sellele järgnev ankeetküsitlus samade hindajate seas. Tulemused näitavad küll kehtestatud hindamisreeglite üldist järgimist mõlemas uurimisrühmas, kuid samas ilmnevad statistiliselt olulised erinevused nii hindajate tegevuses kui ka nende poolt antud hinnangutes kõneoskuse hindamisskaala eri aspektide kohta.

Võtmesõnad: hindamine, hindaja variatiivsus, hindaja usaldusväärsus, kultuuriline valiidsus, suuline keelepädevustest