(Unpublished manuscript)

Rating of Arabic L1 Speakers in a Research Based on the Finnish National Language Proficiency Test

Mia Halonen, JyU

Abstract

This article addresses a change in the rating of spoken language proficiency of Arabic first language (L1) examinees between 2015 and 2016 in a study based on the National Certificate for Language Proficiency (NCLP) test for applying for citizenship in Finland. The study is part of a larger research project *Broken Finnnish: Accent perceptions in societal gatekeeping* (funded by the Academy of Finland 315581) on rating in the NCLP. In the Project data, the speech performances rated in the NCLP 2015 were rated anew in 2016 by 44 raters. The data of this article consist of ratings of the Arabic L1 examinees. As a grade of 3 is the threshold for passing the test, the focus is on the examinees who received a 3 in the NCLP test in 2015 (N = 8) in comparison to their rating a year later in the Project data in 2016. The data thus consist of 352 (8 x 44) ratings. The results showed a drop in the ratings: only 70.40% of the 2015 passers achieved the threshold rating in 2016 as well. Within the male examinees, the passing rate in 2016 was only 53.40 %. The results are discussed in relation to various possible explanatory factors, including the discriminative atmosphere concerning the Arabic L1 migration wave in 2015–2016.

Key words: Citizenship; Language test; Proficiency Rating; Familiarity effects; Arabic L1; Negative bias

Introduction

The National Certificate for Language Proficiency test (NCLP) is a remarkable gatekeeper of citizenship in Finland as the certificate is the most often used means of applying for Finnish citizenship (Ahola 2022: 24). The NCLP test is an audited examination arranged by the Finnish National Agency for Education based on Finnish legislation (Decree 1163; Nationality Act 359). It is a high-stakes test, the results of which in Finnish or Swedish, the official languages of Finland, can be used to apply for citizenship, work, and studies. Among the Arabic L1 participants in focus in this article, almost 90% of the tests are used for citizenship applications.

In this important test there is no room for any kind of bias, and it should be ensured that the test is both fair and just (see e.g., Shohamy 1994, 2001; McNamara 1998.) To study whether there nonetheless might be bias, a research project *Broken Finnnish: Accent perceptions in societal gatekeeping* examining the rating processes of the Finnish language test in the NCLP was designed and implemented within the test setting. Theoretically, the research draws on the critical perspective on language testing (see, e.g., Lynch 2001; Shohamy 2001).

The research data were based on authentic NCLP speech performances and their ratings in the NCLP in 2015, hereafter referred to as the *NCLP*. In the NCLP, professional language raters who are trained in the test system rate the proficiency of the examinees in various language skills. To pass the test, the examinees should receive a grade of 3, which is equivalent to B1 in the Common European Framework of Reference for Languages (CEFR, 2001). For the research project, the same performances were rated by the same raters a year later, in 2016, through data collection on an internet platform, referred to in this article as *reassessment data*. With this design the project aimed at getting knowledge of the general rating process as well as of possible changes in the rating from 2015 to 2016.

In order to study possible biases in the rating, the data was on purpose designed focusing on examinees of five first language (L1) groups, which were known to face negative stereotyping in Finland: Thai, Estonian, Arabic, Russian and Finland Swedish. (For details of the stereotyping, see e.g. Reuter & Kyntäjä 2006; Jaakkola 2009; McRae, Bennett & Miljan 1998; Lumio 2014.) The hypothesis was that if there were bias in the rating, these would be the groups in which it would show. In the NCLP test, the raters receive no background information about the examinees. It is typical, however, that people either consciously or unconsciously infer, right or wrong, speakers' background based on their way of speaking, that is, an 'accent'. For this reason, out of the various language skills rated in the NCLP – listening, speaking, reading, and writing – for the project data, speaking was selected as the focus skill. Speaking is the domain of language use which is immediately available for evaluation of not only the proficiency of the speakers but also of their other features such as their (perceived) gender and age, as well as the linguistic and consequently ethnic background of the speaker and highly subjective impressions of 'likeability'. (See more the Background; also Moyer 2013 for an extensive overview on accent perceptions.)

Previous research has demonstrated that foreign accents are perceived more negatively than native ones, and that language professionals, including test raters, do not differ from naive listeners when it comes to perceptions of accents (e.g. Flege, Munro & MacKay 1995.) In their theory of reversal stereotyping, Kang and Rubin (2009) stated that negative attitudes towards (assumed) speakers can also strengthen the negative perception of accent. Previous research thus suggested that the perceptions of speech are profoundly based on the affective attitudes towards the speaker groups, rather than only on the speech as such (see e.g. Blommaert 2001; Heller 2004). The attitudes can be either positive or negative and might affect the rating.

The monitoring of rating done in the NCLP itself already showed some inconsistent, too severe or too lenient ratings in the test, some of which was rater related, some of which seemed to be based on examinees' L1. Moreover, the research done in the project has shown there are differences between the NCLP ratings in 2015 and the research data ratings in 2016 either in the general rating or in between various criteria when it comes to, for example, Thai, Estonian L1 and Finland Swedish L1 groups (Ahola 2020a, b; Ohranen & Ahola 2022), all leaning towards at least a slightly more lenient rating than expected. When it comes to the Russian L1 examinees, the unpublished statistical analyses (*t* values) showed no significant differences.

Of the five focus groups, the Arabic L1 examinees stood out in that their proficiency was lower in the reassessment data in 2016 than in the NCLP in 2015. In this article I will focus on the Arabic examinees who passed the NCLP in 2015. The research questions explored are as follows: What was the ratio of the Arabic L1 NCLP passers (grade 3) and the Arabic L1 grade 3 examinees in the *Research data* and did the (mis)recognition of the examinee L1 have any effect on the rating in the reassessment data? In the following I present the background of the study, the data and results, and discuss possible reasons for the difference in the rating of the Arabic L1 passers.

Background: The Double-Edged Sword of Familiarity in Proficiency Rating

It is self-evident that raters in any system should aim for justice, fairness, and consistency in their ratings, but for decades it has been a well-researched fact (see e.g. Giles 1970; Derwing & Munro 1997; Lippi-Green 1997; Major et al. 2002) that attitudes towards different groups of people and stereotyping may affect understanding and proficiency ratings. A foreign accent also seems to be something intrinsically recognisable. It has been shown that an accent can be recognised with as little as one word (Purnell, Idsardi & Baugh 1999), by infants (Kinzler 2008), and even in a language the listeners themselves do not know (Major 2007). Even though an accent can easily be recognised as foreign, it can at the same time be considered familiar. It might, for example, be an accent typical for speakers of languages of neighbouring countries, and one might have heard it frequently. A person might also have taught groups of learners of some particular L1s or there might have been a significant amount of media input of people speaking with the accent in question.

Studies have shown repeatedly that familiarity with a speaker's accent improves both understanding and rating of the speech (e.g. Carey, Mannell & Dunn 2011; Fayer & Krasinski 1987). For example, in a study of raters with the same L1 as the speakers under evaluation, Winke, Gass and Myford (2013) demonstrated that this familiarity even led to a positive bias in rating and should thus be monitored in test systems. However, familiarity might as well convey negative associations that will lead towards lower rating. This has been shown, for example, in the attitude studies of Brennan and Brennan (1981) on Mexican-American accents and Cargile's (1997) study on Chinese accents in the USA. Both of the accents were familiar and recognisable for the

American undergraduate listeners who rated both of them very low in proficiency. Lindemann's (2005) study on the non-native English accents of 58 countries and their attractiveness to native English speakers provided further evidence for the argument that familiarity can lead to negative as well as positive bias.

Kang and Rubin (2009) showed that listening can involve so-called reverse linguistic stereotyping. In their experiment, they used Lambert's (1967) matched-guise technique in which listeners hear multiple samples actually spoken by the same speaker, but in which the sample is each time primed with a different hint of the speaker, such as a photo or a name, presenting some easily recognisable stereotype of some ethnicity or language (e.g. an Asian or Arabic name; cf. Ahmad 2020). In Kang and Rubin's research, the listeners heard two speech samples from the same native speaker of English. The sample was primed with a picture of an Anglo-American white man and an Asian man. The listeners with self-reported negative attitudes towards Asian people understood the Asian sample clearly worse than the Anglo-American sample. Thus, attitudes triggered by the priming clearly affected how they heard the sample, as well as their understanding and assessment of the speech. A similar type of reverse stereotyping and hierarchisation of foreign accents, that is, of their speakers, has been reported in Sato's (1998; presented in Munro 2003) study of foreign-accented English in Australia. In it, Australian listeners were to rate speakers' personality. In the case of a Ukrainian L1 speaker, some of the raters misperceived the speaker being of Aboriginal background and assessed the speaker more negatively than the ones correctly identified as Ukrainian L1 speakers. When raters assume that they have heard a speaker of a specific language, they rate the speaker also as a representative of that language group, however wrong their assumptions might be. A rating can therefore be affected by stereotypes and attitudes connected to the assumed group.

In many real-life situations, for example in the NCLP and in the research in focus of this article, the performances to be rated are not primed in any way and the raters receive no background information of the speakers. As the abovementioned research shows, especially Sato's, this does not inevitably lead to a neutral stance. Rather, the raters might make assumptions, consciously or unconsciously, about an examinee's background based on clues offered by the performance itself. Pantos and Perkins (2012) have shown that people's explicit attitudes can differ completely from their implicit attitudes when it comes to foreign accents. Similarly, also (meta)linguistic awareness and knowledge, as well as the ability to reflect on them, vary hugely even among language professionals (Preston 1996). Professionals, including language teachers and language raters, can be as prejudiced as any so called naive listeners. This being the case, biased rating is possible also among the experienced and educated NCLP raters.

What the previous research on various aspects of speakers' assumed or recognised background affecting rating maintain is that familiarity and (mis)recognition of the speakers' L1 is a complex double-edged phenomenon in which attitudes and stereotypes connected to the 'familiar' play a role and can lead to both positive and negative perceptions and biased ratings. A foreign accent in general can be problematic, not recognising the L1 can be problematic, recognition of the L1 might turn out to be even more problematic if the L1 is recognised as conveying negative stereotypes, and it is at least as problematic if the L1 is recognised wrongly and the speaker will be rated according to the wrong assumption. Indeed, in the case of rating speakers with a foreign accent, there is no escape from potential problems.

To complicate the possible effects of familiarity even further, raters may also evaluate speakers on behalf of others. This has been shown to happen in, for example, employment interviews (e.g. Deprez-Sims & Morris 2013; Kokkonen 2007). The interviewers might find the applicant's accent easy to understand but are sceptical when it comes to their customers' possible reaction towards it. Thus, they might evaluate the proficiency in relation to their perception of what others in the society require from a representative of that particular profession. Also in the context of the research reported here, the raters are not only professional raters but also members of Finnish society. It is possible that society's attitudinal atmosphere affects the rating or that the raters take the (assumed) societal atmosphere into account when rating. For this reason, I will next present a glimpse of the societal atmosphere in Finland in 2015–2016 as the context of the study.

Context of the study: Arabic L1 speaker migration wave and the societal atmosphere

In 2015–2016, that is, the time span between the NCLP rating and when the research data was collected, Finland experienced a migration wave of Arabic L1 speakers, and this resulted in a general attitudinal atmosphere created by or associated with that wave.

The so-called Arab Spring, a series of protests and armed rebellions in Northern Africa and the Middle East, starting from the revolution in Tunisia in 2010 and leading to serious conflicts and civil wars, including the 2012 Syrian civil war, has been a huge global humanitarian crisis. This was and still is manifested in the massive refugee and migration wave especially in Southern Europe but as far north as Finland, where the peak of the wave took place in 2015–2016. Statistics show that in Finland the relative growth of Arabic L1 speakers from 2015 to 2016 was 30.34%. Even though the numbers of Arabic L1 speaking migrants have been more than modest in absolute measures, being 12,042 in 2015 and 21,783 in 2016, the relative growth can be described as remarkable.

This growth had various consequences. First, relevant to the study at hand, it led to a rise in raters' encounters with Arabic L1 speakers, both as the NCLP rater examiners and as teachers of Finnish as a second language (L2). This has naturally also raised their familiarity with the Arabic accent in Finnish. Even though in this migration population there were speakers of various L1s and various dialects and varieties of Arabic, a clear majority reported Arabic as their L1 in bureaucratic contexts, such as in the NCLP.

The growth also affected the attitudinal atmosphere. In the European Union's surveys on discrimination in its Member States (EU-MIDIS 2009, 2018), Finland has repeatedly been revealed to be one of the most discriminating and racist countries in Europe. According to the EU-MIDI 2017, which focused on the Muslim population, everywhere in Europe more or less half of Muslims reported having faced discrimination in housing, job applications and healthcare based on their first and last names. When it comes to Finland, Pauha and Ketola (2015) showed, based on a 2011 poll about attitudes towards religions, that Finnish attitudes towards Islam are particularly negative even in international comparison. A fairly recent study by Ahmad (2020) again showed similar results: an Arabic name was a huge disadvantage in the labour market in Finland.

Arabic, the language of the Qur'an and therefore a lingua franca of Muslims in general, has an integral relation with Islam and, consequently, with Islamophobia stemming especially from associations with violent terrorism. (See e.g. Allen 2010; for more about Islamophobia in Finland, see e.g. Martikainen 2020). According to Iqbal (2019), language forms one of the epistemic dimensions of Islamophobia, along with religious, social, and cultural components. In current Western societies, *Arabic L1 speaker* is in fact a categorisation which stands at an intersection of language, religion, and ethnicity, along the lines of Crenshaw's (1991) definition of political intersectionality. The categorisation is based on both the real but also imagined exclusive relation of the Arabic language and Islam.

In the context of Finland, the migration wave of 2015–2016 seemingly even strengthened phenomena such as Islamophobia and discrimination, drawing a significant amount of pro and con activism while attracting a huge amount of mass media and social media attention (e.g. Laaksonen, Pantti & Titley 2020; Maasilta & Nikunen 2018), both reflecting and constructing the general attitudinal atmosphere. Along with the overall rise in the encounters of raters and Arabic L1 speakers, the atmosphere might have played a role in the rating of the Arabic L1 examinees.

Data and Methods

The growing amount of data produced as part of the NCLP test are not open or public for obvious reasons. First, the data are sensitive and confidential, the examinees being in a vulnerable position. Second, legislation as well as personal data register regulations (e.g. GDPR, IT Governance Privacy Team) set limits on what data can be used and how. The strict confidentiality of the NCLP was one of the reasons for creating a separate research project focusing on the test system.

In the NCLP test, the examinees complete the test at test organising institutions which are located everywhere in Finland, mostly concentrated in the bigger cities. The performances are then assessed in a centralised rating session where professional raters from all over the country come together to rate them. When taking the test, the examinees give their written consent for the use of the performance in the NCLP and in studies conducted within the test system, such as the one reported on here. The raters, too, have given permission for the rating data to be used in research. In the research reports, all the participant information is anonymised. Because the data are both sensitive and the confidentiality regulated by legislation, they are not publicly available. The metadata of the data are available in JyX archive (https://doi.org/10.17011/jyx/dataset/85233).

In the Finnish language test, there are altogether slightly over 100 raters. Typically, there have been five test sessions a year. The raters are language professionals, most often teachers of Finnish as a second language (L2). They have been trained for the NCLP rating and their severity and consistency is monitored in every rating session. There is also a joint training session at the beginning of each rating session, in which the raters study and discuss together some cases as examples of various proficiency levels. The proficiency level required for citizenship applications is a rating of 3 on a scale of 1–6, which is equivalent to B1 level in the Common European Framework of Reference for Languages (CEFR 2001). In the NCLP, the raters give only one holistic proficiency grade for the performance.

The reassessment data for the project were built on the basis of the actual results of the NCLP, more precisely on the examinee performances and the ratings that were carried out in 2015. For the reassessment, 44 raters who had agreed to participate in the research rated the performances anew in 2016 on an internet-based platform. In the reassessment, 10 speech performances of each of the focus groups of Thai, Estonian, Finland Swedish, Arabic and Russian speakers (5 from two binary genders of men and women indicated by the examinees themselves in a background questionnaire) were rated. The performances included both performances of examinees who had passed the test in 2015, that is, received a minimum grade of 3, as well as performances that did not pass, that is, received a grade under 3.

In the reassessment, in addition to the holistic proficiency rate, the raters also rated the performances in relation to six analytical criteria of fluency, flexibility, coherence, vocabulary, pronunciation, and grammatical accuracy, on a scale of 1–6. In addition to these ratings, they described the examinees and their performances and gave an assumption of the examinees' first language, for the researchers to know who the raters think they were rating, and provided their reasoning for the assumption as an open answer. As mentioned earlier, in the NCLP test, raters receive no background information of the examinees and indicate in no way whether they have recognised, assumed, or even thought about the examinees' background.

In the reassessment the design was similar in that the raters had no information about the examinees' background. However, for the purposes of the research, the raters were asked to make an assumption of the examinees' L1 and give a reasoning for the assumption. This was done because, when studying attitudes towards languages and speakers, it is crucial to know who the raters think they heard, which has been pointed out by language attitude researchers Preston and Niedzielski (e.g. 2013). That the raters do not do this normally possibly showed in that in 12% of cases this question was not answered or received answers such as 'I don't know' or 'I don't want to guess'. Nevertheless, it is clear that for the majority of the raters (78%), assumption of L1seems to have been an implementable and even easy task. In 31.4% of the total cases (*N* = 2,159), assumption of L1 was answered first, even though it was only the tenth out of twelve questions or rating items.

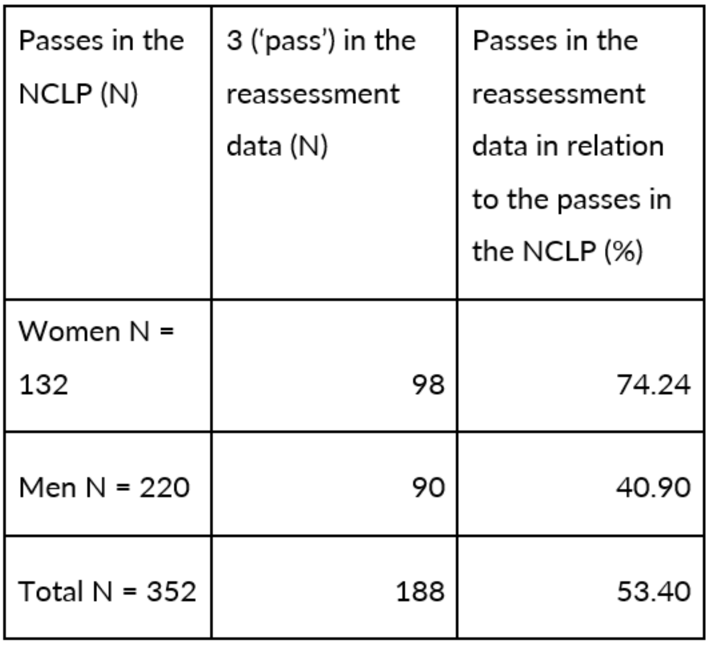
Because the reassessment in 2016 were built on the NCLP from 2015, the data enable comparison of the grades In addition, the reassessment data as a whole are extensive, in the research reported in this article, the focus has been narrowed to crucial, real-life threshold of *pass* or *not pass*. Thus, only the Arabic L1 examinee performances that passed in the NCLP (received 3 for the overall grade) have been studied and the grades for 2015 and 2016 were compared. In the case of Arabic L1 examinees in the reassessment data, 8 of 10 of them were passers in the NCLP.

All 5 male examinees in the reassessment data had passed in the NCLP, as well as 3 of the 5 female examinees. Since there were 44 raters, there were 352 NCLP pass cases among the Arabic L1 speakers: 220 men (5 x 44) and 132 women (3 x 44). In the analyses, the quantity of the NCLP passes was compared to the quantity of the examinees who received a grade of 3 in the reassessment as well. Hence, the main method used was simply descriptive statistics comparing the relative quantity of passes in the NCLP from 2015 and in the reassessment data from 2016. In the reassessment data, the assumptions of L1 were also explored in relation to the passing grades.

Results

As presented above, to pass the NCLP test as well as in the research data, the examinees needed to receive a grade of 3. Table 1 shows the quantity and ratio of passing grades in the NCLP and in the reassessment data. It also presents the female and male examinees separately. In the case of the Arabic L1 speaker examinees, gender happens to be binary, male and female, chosen by the examinees themselves in a background inquiry. The raters were not asked to assume any gender of the examinees. However, I maintain that the genders were perceived similarly by the raters as they were by the examinees as well as by the researcher. All the samples are prototypical when it comes to the most typical physical differences of female and male vocal tracts that lead to universal differences in formant frequencies, for example in pitch, which again often leads to the perception of a speaker as either man or woman (see e.g. Fant 1966.)

Table 1. Passing grades in the reassessment data in relation to the NCLP rating.

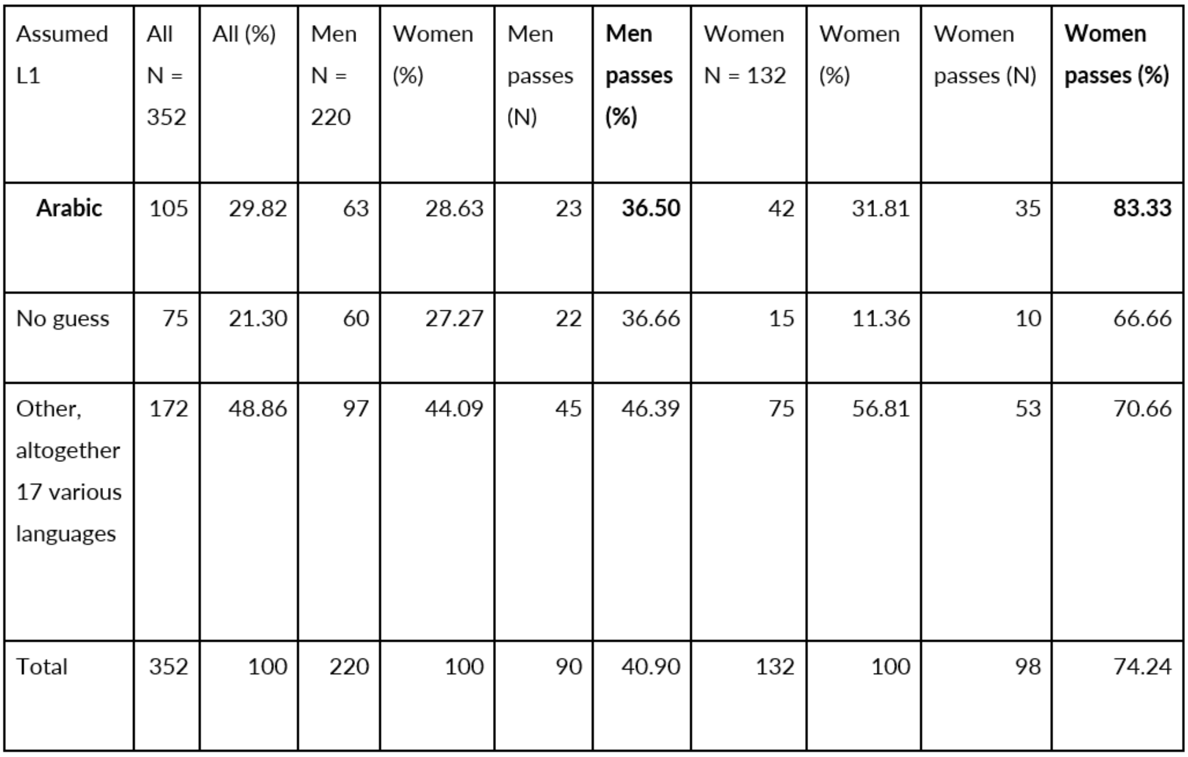


Passes in the NCLP (N)	3 ('pass') in the reassessment data (N)	Passes in the reassessment data in relation to the passes in the NCLP (%)
Women N = 132	98	74.24
Men N = 220	90	40.90
Total N = 352	188	53.40

Since the performances rated were exactly the same at both points, it was expected, based on the NCLP, that all those who passed the NCLP would have passed in the reassessment as well. However, this was not the case, as only 53.40% of those who passed in the NCLP received a grade of 3 in the reassessment. Although both men and women were rated lower than expected, the table also shows a huge difference between the relative quantity of passers of men (40.90%) and women (74.24%).

As argued above, to be able to study potential attitudes affecting rating, it was crucial to know who the raters thought they were rating, thus, what they assumed to be the L1 of the speakers. As the focus was on Arabic L1 speakers, the most important factor to know was the degree of Arabic L1 assumptions, that is, correct recognition. Table 2 presents the assumptions and their distribution together with the gender distribution. The main findings that will be discussed below have been bolded.

Table 2. Assumed L1s within Arabic L1 examinees in relation to the passes in the NCLP



Assumed L1	All N = 352	All (%)	Men N = 220	Women (%)	Men passes (N)	**Men passes (%)**	Women N = 132	Women (%)	Women passes (N)	**Women passes (%)**
Arabic	105	29.82	63	28.63	23	**36.50**	42	31.81	35	**83.33**
No guess	75	21.30	60	27.27	22	36.66	15	11.36	10	66.66
Other, altogether 17 various languages	172	48.86	97	44.09	45	46.39	75	56.81	53	70.66
Total	352	100	220	100	90	40.90	132	100	98	74.24

In general, the Arabic L1 speakers were not easily recognised. However, Arabic was still the most often assumed L1 of the examinees (29.82%) for men (28.63%) as well as for women (31.81%). The next biggest group were the 'no guess' answers, that is, no answer was given for the L1 (all 21.30%; men 36.66%; women 11.36%). Here the difference between men and women was large: women's L1s were assumed, whether right or wrong, more than twice as often as the men's were. For the rest of the L1 assumptions, 17 different languages were guessed. Of these, no single individual language was assumed more than a maximum of 14 times. That is why they are here reported as a cluster of 17 languages.

The next biggest group, the 'no guess' answers, was quite a remarkable category in numbers. They varied from not answering anything to descriptions of why an L1 assumption was not given. Some stated plainly that they 'did not know' or 'did not want to guess', while some seemed to resist the very task by not responding or giving any explanation as to why they did not respond. In the NCLP, where the background of the examinees is not addressed or discussed at any phase, this is the normal practice, but in the reassessment, where an L1 assumption was specifically asked for, leaving it unanswered is also a statement. While not paying attention to any hint of the examinees' background can be seen as a fair act, it may also be a way of hiding biases. In fact, there were a couple of raters who declined to answer this question and who were clearly biased in relation to one or more L1 groups in the data, as revealed by their rating behaviour.

The results are clear: the Arabic L1 speakers did not reach the expected 100% rate of passing in any of the L1 assumption cases. However, there were also huge differences between the rating of genders. That was the case throughout the reassessment data but was highlighted when it came to the recognition of Arabic as the speakers' L1. Arabic L1 men passed most rarely when the L1 was correctly recognised, in only 36.50% of the cases. The difference between the 2015 and 2016 rating of the male examinees' performances was ample. All the other assumptions, 'no guess' as well as all the other assumed L1s, were favourable when compared to Arabic. At the same time, women passed most often when they were recognised as Arabic L1 speakers, in 83.33% of the cases. They benefited from the correct recognition. The difference between genders was so remarkable that it threatens to blur the fact that they also failed to achieve the expected 100% rate of passing.

The results showed there was a clear decrease in the amount of those who received a grade of 3 from NCLP rating in 2015 and the reassessment in 2016. This applied to all the Arabic L1 speakers in the data (N = 8, rated by 44 raters), but was especially strong for the male examinees who were correctly recognised as Arabic L1 speakers.

At the same time, analyses of the NCLP data from the regular testing rounds (2012–2019) show that during the last ten years, the raters have become more lenient in their ratings as a whole (YY, under review). Considering this otherwise increasing leniency, the increased severity towards Arabic L1 examinees in the one-year period from 2015 to 2016 appears even harsher. This results suggests there may be negative bias against Arabic L1 speakers, which I will discuss next.

Discussion

The decrease from 2015 to 2016 in the amount of Arabic L1 examinees receiving a passing grade is significant in its proportion and especially in its potential consequences. If, for any reason, the rates of passing the test decrease, it would directly and negatively affect the lives of the examinees, their integration into society, working life and studies. The data collected at the reassessment fails to provide direct explanations as to why the decrease took place. Some arguments or similarities in cases can be found in previous research, in the other results and analyses of in the project, and in the surrounding societal context.

This type of data based on an actual language test system has not been studied previously. All the possible reasons are thus somewhat speculative and numerous. An obvious but possible and important one is that rating a performance in an actual NCLP test which gives the examinee an official result is profoundly different from rating the performances in a study that has no affect on the lives of the examinees in any way and has no real-life consequences. Without these consequences in their mind, raters may have felt freer to take a more severe approach in their rating. Moreover, even though the authentic speech performances of the NCLP were used for the reassessment, there was different and additional information gathered compared to the NCLP test situation, for example the assumed L1. With this, the reassessment data gathering unavoidably directed the raters' attention to the examinees themselves and their background instead of only the performance. Even if that were the 'simple' answer to the drop, it does not mean the result is trivial or insignificant. Instead, if that were the explanation for the drop, it would mean that in the actual NCLP the raters need to monitor their rating and act concentratedly and consistently in favour of the examinees.

A further argument against the decrease having only been about the difference between test and research contexts is that there was no such decrease—or none at all—in the other L1 groups. Of special interest is the difference between the rating of Thai L1 and Arabic L1 examinees. According to the NCLP internal analyses, these L1 groups have very similar so-called general ability, which means they receive approximately similar grades in the NCLP. However, in the reassessment, the Thai L1 examinees were rated similarly as they were in the NCLP. What might also play a role in the rating are the L1 assumptions and skill expectations associated with them: expectations of low skills might lead to higher ratings, and, vice versa, high expectations to lower ratings. (See Lindemann 2006 for an overview on various studies on the effects of expectations.) Expectation wise, thus, it would be expected that Thai and Arabic L1 examinees would be rated similarly in the reassessment data. That was not the case, but the Arabic L1 examinees were rated lower.

The results also clearly showed that the men were much more severely rated than women were. One factor in this difference could be the gender difference between the raters and the examinees. In the reassessment data, there were 4 male and 40 female raters, which equals the general ratio of men and women in NCLP raters. Gender differences between raters and examinees have rarely been studied. However, Buckingham's (1997) study on this effect showed a general leniency towards the same gender. That might partly also explain the difference between the rating of Arabic and Thai L1 examinees. This possibility is also supported by Ahola (2020a) on Thai L1 examinees, which shows, based on analyses of the open answer explanations, that the raters seem to have a positive, compassionate, even patronising attitude towards the Thai women examinees, and that they were lenient towards them. When it comes to general attitudes, such as societal discrimination, gender seems to matter. In general, outside language test contexts, there seems to be negative bias towards male L2 speakers (see e.g. Gallois & Callan 1981). Furthermore, if the association between Arabic and Islam has played a role in the rating, EU-MIDIS (2017) showed that Muslim men face even more discrimination than Muslim women do.

In the research data, Arabic L1 examinees were poorly recognised, but Arabic was nevertheless their most commonly assumed L1. The familiarity with the Arabic L1 accent had most probably increased due to the migration wave. As discussed in the background, familiarity can have negative as well as positive effects. An extensive amount of research has shown that familiarity with speakers' accents improves both comprehension and rating. In this study, however, recognition of the Arabic L1, stemming most probably from familiarity, had a negative effect on the rating. Drawing on that, it could then be expected that, consequently, that unfamiliarity, that is the raters not having recognised the Arabic L1 examinees, should have led to higher ratings. This did not happen either, which then again, aligns with the previous research suggesting only a rise in rating along with familiarity. In the case of the Arabic L1 examinees, neither familiarity nor unfamiliarity worked in their favour. The results fit anyway into the overall picture of familiarity, as it genuinely is a double-edged sword. It can increase intelligibility because it makes speakers easier to understand, but familiarity can also carry unfavourable associations, again possibly affecting the rating negatively (cf. Derwing & Munro 1997).

The size of the drop in those who received a passing grade as well as the difference of the rating in the reassessment compared to other L1 focus groups, especially to Thai speakers, suggest negative bias against the Arabic L1 examinees, especially towards men. Again, just like all the possible explanations for the drop, the bias can be based on various sources. The raters might themselves have negative experiences of Arabic L1 speakers as, for example, their students, or they might, for the same reason, have higher expectations for their proficiency and consequently, rate the performances lower in light of them. It is also possible that the general societal atmosphere in Finland at the height of Arabic L1 migration during 2015–2016, and the escalated polarisation of attitudes for and against migration presented in the context of the study had an effect. It could have made the raters' own attitudes towards Arabic L1 speakers harsher. It is possible that they 'rate for others' (cf. e.g. Kokkonen presented in the background), making them more severe in their rating in order to correspond to the surrounding expectations, not letting examinees pass with skills that are too low in a hostile society.

Having discussed all this, it is necessary to further consider whether the results and their speculative explanations are reliable or convincing. It could be argued that we should have had another type of data, such as interviews with the raters. However, it is unlikely that we could obtain any reliable data on such sensitive topics as attitudes and biases through interviews. In most situations, people typically present themselves in a positive light. This is particularly true for professional raters who are expected to be especially fair and unbiased. With the type of data in this study, which was designed to provide information about various aspects of the NCLP rating process and in which bias and attitudes are studied indirectly through the rating behaviour itself, all explanations remain speculative. Yet this does affect the results, which should be discussed and have an impact on the test system.

Conclusion

The empirical focus in this article was on the change in rating of speech proficiency ratings of Arabic L1 examinees in the NCLP in 2015 and the reassessment in 2016. The results showed that there is a negative bias towards Arabic L1 examinees, and especially towards men. Although with this data it is impossible to know exactly why this is the case, other research on attitudes towards L2 speakers and examinees, and specifically on attitudes towards Arabic L1 speakers, point to some type of a negative attitude towards them leading to lower grades than those predicted by the NCLP rating. Previous research on attitudes affecting rating and results presented in this article suggest that even the NCLP qualified professional raters might be affected by their own attitudes or those from the surrounding society.

The NCLP is a high-stakes test with serious consequences. For its organisers, who come from within the system, the negative bias shown in the reassessment data is disconcerting to say the least, and we need to look at this result as it is. We intend, through research, to take part in language test activism, drawing on critically oriented pioneering scholars who aim to promote justice in language tests, such as Shohamy (e.g. 1994) and McNamara (e.g. 1998). Although a critical approach questions the entire ideologically and politically loaded institution of language testing, it is also possible to take a critical stance from an insider position within a test system and, through research, to point out potential problems and their causes, with the aim of diminishing them. High-stakes tests must be open for development through academic research.

By empirically studying the rating in the system, we aim to provide tools for promoting equality. One of the steps is to return the research results to the system and include knowledge of possible conscious or unconscious attitude- or stereotype-based biases in the rater education alongside with the monitoring of general rater behaviour of severity and consistency. The results suggest that, for example, increasing familiarity with an L1 group and its accents is not efficient or enough to control and correct biases. As we have seen, familiarity can be a double-edged factor which can lead to both positive and negative bias. The most efficient way of ensuring fair ratings is to educate the raters about bias, attitudes, and stereotyping.

References

Ahmad, A. 2020. When the name matters: An experimental investigation of ethnic discrimination in the Finnish labor market. *Sociological Inquiry*, 90(3): 468–496. https://doi.org/10.1111/soin.12276

Ahola, S. 2020a. Sujuvaa mutta viron kielen vaikutusta. *Virittäjä* 124 (2), 217–242. https://doi.org/10.23982/vir.79831

Ahola, S. 2020b. Yleisten kielitutkintojen arvioijien käsityksiä thainkieliseksi tunnistettujen suomenoppijoiden suullisesta kielitaidosta. *Puhe ja kieli* 40 (4): 203–224. https://doi.org/10.23997/pk.103307

Allen, C. 2010. *Islamophobia*. Farnham: Ashgate.

Blommaert, J. 2001. The Asmara Declaration as sociolinguistic problem: reflections on scholarship and linguistic rights. *Journal of Sociolinguistics*, 5: 131–142. https://doi.org/10.1111/1467-9481.00142

Brennan, E. & Brennan, J. S. 1981. Measurements of accent and attitude toward Mexican-American speech. *Journal of Psycholinguistic Research*, 10: 487–501. https://doi.org/10.1007/BF01076735

Buckingham, A. 1997. Oral language testing: do the age, status and gender of the interlocutor make a difference? Unpublished MA dissertation, University of Reading.

Carey, M. D., Mannell, R. H. & Dunn, P. K. 2011. Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, 28(2): 201–219. https://doi.org/10.1177/0265532210393704

Cargile, A. C. 1997. Attitudes toward Chinese-accented speech: An investigation in two contexts. *Journal of Language and Social Psychology*, 16(4): 434–443. https://doi.org/10.1177/0261927X970164004

CEFR 2001 = *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe.

Crenshaw, K. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6): 1241–1299. https://doi.org/10.2307/1229039

Decree on the National Proficiency Certificate of Language Proficiency 1163/2004. Asetus yleisistä kielitutkinnoista. https://www.finlex.fi/fi/laki/ajantasa/2004/20041163
[Last accessed 5 January 2023].

Deprez-Sims, A.-S. & Morris, S. B. 2013. The effect of non-native accents on the evaluation of applicants during an employment interview: The development of a path model. *International Journal of Selection and Assessment*, 21(4): 355–367. https://doi.org/10.1111/ijsa.12045

Derwing, T. & Munro, M. 1997. Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition*, 19(1): 1–16. https://doi.org/10.1017/S0272263197001010

EU-MIDIS 2009. *European Union Minorities and Discrimination Survey. Main Results Report*. Luxembourg: Publications office of the European Union. https://fra.europa.eu/sites/default/files/fra_uploads/663-fra-2011_eu_midis_en.pdf [Last accessed 11 January 2023].

EU-MIDIS 2017. *Second European Union Minorities and Discrimination Survey. Muslims – Selected Findings*. Luxembourg: Publications office of the European Union. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2017-eu-minorities-survey-muslims-selected-findings_en.pdf [Last accessed 11 January 2023].

EU-MIDIS 2018. *Second European Union Minorities and Discrimination Survey. Being Black in the EU*. Luxembourg: Publications office of the European Union. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-being-black-in-the-eu-summary_en.pdf [Last accessed 11 January 2023].

Fant, G. 1966. A note on vocal tract size factors and non-uniform F-pattern scalings. *STL-QPSR*, 4: 22–30.

Fayer, J. M. & Krasinski, E. 1987. Native and nonnative judgments of intelligibility and irritation. *Language Learning*, 37(3): 313–326. https://doi.org/10.1111/j.1467-1770.1987.tb00573.x

Flege, J., Munro, M., & MacKay, I. 1995. Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, 97: 3125–34. https://doi.org/10.1121/1.413041

Gallois, C. & Callan, V. 1981. Personality impressions elicited by accented English speech. *Journal of Cross-Cultural Psychology*, 12(3): 347–359. https://doi.org/10.1177/0022022181123006

Giles, H. 1970. Evaluative reactions to accents. *Educational Review*, 22(3): 211–227. https://doi.org/10.1080/0013191700220301

Heller, M. 2004. Analysis and stance regarding language and social justice. In: D. Patrick & J. Freeland *Language rights and language survival: Sociolinguistic and sociocultural perspectives*. Manchester: Jerome, pp. 283–286.

Iqbal, Z. 2019. *Islamophobia: History, Context and Deconstruction*. Los Angeles: Sage.

IT Governance Privacy Team. *EU General Data Protection Regulation (GDPR) – An Implementation and Compliance Guide, Fourth Edition*. Vol Fourth edition. ITGP; 2020.

Jaakkola, M. 2009. *Maahanmuuttajat suomalaisten näkökulmasta: Asennemuutokset 1987–2007*. Helsinki: Helsingin kaupungin tietokeskus.

Kang, O. & Rubin, D. 2009. Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, 28(4): 441–456. https://doi.org/10.1177/0261927X09341950

Kinzler, K. 2008. The native language of social cognition: Developmental origins of social preferences based on language. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Kokkonen, M. 2007. Vaatimuksena sujuva suomi. *Virittäjä*, 111: 253–261.

Laaksonen, S., Pantti, M., & Titley, G. 2020. Broadcasting the movement and branding political microcelebrities: Finnish anti-immigration video practices on YouTube. *Journal of Communication*, 70(2): 171–194. https://doi.org/10.1093/joc/jqz051

Lambert, W. E. 1967. A social psychology of bilingualism. *Journal of Social Issues* 23(2): 91–109. https://doi.org/10.1111/j.1540-4560.1967.tb00578.x

Lindemann, S. 2005. Who speaks `broken English´? US undergraduates' perceptions of non-native English. *International Journal of Applied Linguistics*, 15(2): 187–212. https://doi.org/10.1111/j.1473-4192.2005.00087.x

Lindemann, S. 2006. What the other half gives: The interlocutor's role in non-native speaker performance. In: Hughes, R *Spoken English, Tesol and Applied Linguistics. Challenges for Theory and Practice*. London: Palgrave MacMillan. pp. 23-49. https://doi.org/10.1057/9780230584587_2

Lippi-Green, R. 1997. *English with an accent: Language, ideology, and discrimination in the United States*. New York: Routledge.

Lumio, M. 2014. Hymyn takana – thainmaalaiset maahanmuuttajat ja suomalais-thainmaalaiset avioliitot. In: Heikkilä, O. & Säävälä, M. *Monikulttuuriset avioliitot sillanrakentajina*. Turku: Migration Institute of Finland. pp. 36−51.

Lynch, B. K. 2001. Rethinking assessment from a critical perspective. *Language Testing*, 18(4): 351–372. https://doi.org/10.1177/026553220101800403

Maasilta, M. & Nikunen, K. (eds.) 2018. *Pakolaisuus, tunteet ja media*. Tampere: Vastapaino.

Major, R. 2007. Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition*, 29(4): 539–556. https://doi.org/10.1017/s0272263107070428

Major, R., Fitzmaurice, S., Bunta, F. & Balasubramanian, C. 2002. The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, 36(2): 173–190. https://doi.org/10.2307/3588329

Martikainen, T. 2020. Finnish Muslims' journey from an invisible minority to public partnerships. *Temenos Nordic Journal of Comparative Religion* 56(1): 33–51. https://doi.org/10.33356/temenos.77424

McNamara, T. F. 1998. Policy and social considerations in language assessment. *Annual Review of Applied Linguistics* 18: 304–319. https://doi.org/10.1017/S0267190500003603

McRae, K.D., S.E. Bennett & Miljan, T. 1988. *Intergroup sympathies and language Patterns in Finland: Results from a survey*. Helsinki: Suomen Gallup.

Moyer, A. 2013. *Foreign accent – the phenomenon of non-native speech*. Cambridge: Cambridge University Press.

Munro, M. 2003. A primer on accent discrimination in the Canadian context. *TESL Canada Journal*, 20: 38–51. https://doi.org/10.18806/tesl.v20i2.947

Nationality act 359. Kansalaisuuslaki. https://finlex.fi/fi/laki/ajantasa/2003/20030359 [Last accessed 5 January 2023].

Ohranen, S. & Ahola, S. 2022. "Vierasta korostusta on niin vähän, että arvaan viro": yleisten kielitutkintojen suomen kielen arvioijien käsityksiä suomenruotsalaisten puhumasta suomesta. *Lähivõrdlusi-lähivertailuja*, *32*, 150–180. https://doi.org/10.5128/lv32.05

OSF 2020 = Official Statistics of Finland. *Population structure* [e-publication].
https://www.stat.fi/til/vaerak/index_en.html
[Last accessed 5 January 2023].

Pantos, A. J. & Perkins, A. W. 2012. Measuring implicit and explicit attitudes toward foreign accented speech. *Journal of Language and Social Psychology*, 32(1): 3–20. https://doi.org/10.1177/0261927X12463005

Pauha, T. J. & Ketola, K. H. 2015. Mikä selittää suomalaisten islam-vastaisuutta? In: Hämäläinen, R & H. Pesonen *Kohtaamisia: Kirjoituksia uskonnosta, arjesta ja monikulttuurisuudesta*. Helsinki: Helsingin yliopisto. pp. 94–105.

Preston, D. R. 1996. Whaddayaknow – modes of folk linguistic awareness. *Language Awareness* 5(1): 40–74. https://doi.org/10.1080/09658416.1996.9959890

Preston, D. R. & Niedzielski, N. 2013. Approaches to the study of language regard. In: Kristiansen, T. & S. Grondelaers *Language (De)standardization in late Modern Europe: Experimental Studies*. Oslo: Novum. pp. 287–306.

Purnell, T., Idsardi, W., & Baugh, J. 1999. Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology*, 18(1): 10–30. https://doi.org/10.1177/0261927X99018001002

Reuter, A. M., & Kyntäjä, E. 2006. Kansainvälinen avioliitto ja stigma. In Martikainen, T. *Ylirajainen kulttuuri: Etnisyys Suomessa 2000-luvulla.* Helsinki: Finnish Literature Society. pp. 104–125.

Sato, K. 1998. Evaluative reactions towards 'foreign accented' English speech: The effects of listeners' experience on their judgements. Unpublished master's thesis, University of Alberta.

Shohamy, E. 1994. The use of language tests for power and control. In: Alatis, J. *Georgetown University round table on language and linguistics*. Washington DC, Georgetown University Press. pp 57–72.

Shohamy, E. 2001. *The power of tests: A critical perspective on the uses of language tests*. Harlow: Pearson Education.

Winke, P., Gass, S., & Myford, C. 2013. Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30(2): 231–252. https://doi.org/10.1177/0265532212456968