Snap Judgment : Influences of Ethnicity on Evaluations of Foreign Language Speaking Proficiency

Claire Gilchrist, Jean-Pierre Chevrot

Corela
Cognition, représentation, langage

15-1 | 2017
Vol.15, n°1

Snap Judgment : Influences of Ethnicity on Evaluations of Foreign Language Speaking Proficiency

Claire Gilchrist et Jean-Pierre Chevrot

Édition électronique
URL : http://journals.openedition.org/corela/4920
DOI : 10.4000/corela.4920
ISSN : 1638-573X

Éditeur
Cercle linguistique du Centre et de l'Ouest - CerLICO

Référence électronique
Claire Gilchrist et Jean-Pierre Chevrot, « Snap Judgment : Influences of Ethnicity on Evaluations of Foreign Language Speaking Proficiency », *Corela* [En ligne], 15-1 | 2017, mis en ligne le 26 juillet 2017, consulté le 23 novembre 2018. URL : http://journals.openedition.org/corela/4920 ; DOI : 10.4000/corela.4920

Snap Judgment : Influences of Ethnicity on Evaluations of Foreign Language Speaking Proficiency

Claire Gilchrist et Jean-Pierre Chevrot

1. Language Proficiency Tests

1 Language proficiency tests have important consequences for test takers and for society, acting as "gatekeepers" for education, employment, and citizenship (Lippi-Green 2012 : passim ; Jenkins & Parra, 2003 : 90). It is important to ensure that judgments on these high-stakes tests are reliable and valid. Traditional methods of determining rater validity measure consistency between different raters. While these methods can ensure internal validity, they cannot detect biases that are shared across a group of raters (Lindemann & Subtirelu 2013 : 584).

2 Researchers are particularly concerned with the validity of speaking tasks because of the nonlinguistic information conveyed in person-to-person interviews that may affect judgments (Nguyen 1993 : 1335 ; Kang 2008 : 18). Ensuring validity on these tasks is hampered by the qualitative and complex nature of the scoring process. In the TOEFL speaking test, for example, after several minutes of conversation, raters must determine if the test taker's "language use" is a level 1, 2, 3, or 4. The descriptions of the difference between these levels is nuanced ; a level 4 is "effective use of grammar and vocabulary" whereas a level 3 is "*fairly automatic* and effective use of grammar and vocabulary" (Educational Testing Service, emphasis added for clarity).

3 Studies from the field of speech perception have demonstrated the inherent subjectivity of listening. Many experiments have demonstrated the effect of stereotypes on perception at the

4 phonetic level (e.g. Hay & Drager 2010 : passim ; Rubin, 1992 : passim) ; however, none have investigated how these stereotypes might also interfere at the word or sentence level, specifically affecting a listener's ability to "hear" specific grammar or syntax errors.

This question has serious implications for high-stakes language testing. In the present study, we addressed this gap by exploring whether the perceived ethnicity of a speaker could affect the number of grammar errors detected, and whether this quantitative measure is related to final qualitative scores assigned.

1.1. Linguistic Stereotyping and Reverse Linguistic Stereotyping

5 Speech perception can be influenced by stereotypes a listener has about the social group they believe the speaker belongs to. These stereotypes can be activated in just a few seconds of hearing a particular speech pattern, but they can also be activated by beliefs and expectations a listener has about the speaker before they even begin to speak. These two processes are referred to as *linguistic stereotyping* (LS), which denotes phonetically-triggered stereotyping, and *reverse linguistic stereotyping* (RLS), which denotes expectation-based stereotyping.

6 The LS hypothesis posits that positive or negative stereotypes are activated in listeners when they hear a particular speech pattern (Bradac, Cargile & Halleet 2001 : passim ; Rubin 2012 :12). A classic study demonstrating this phenomenon was conducted by Lambert and his colleagues in the 60's (Lambert et al., 1960 : passim). In their experiment, several bilingual Québécois men recorded two speech samples : a passage in French and the same passage translated into English. Next, experimenters asked bilingual participants to listen to the clips and rate each speaker on traits such as attractiveness, sense of humour, and dependability. Ratings of a speaker differed depending on what language he was speaking ; participants demonstrated a tendency to rate the French guise more negatively than the English guise of the same speaker.

7 Recent studies have highlighted the existence of a different type of linguistic stereotyping, called reverse linguistic stereotyping (Kang & Rubin 2009 : 442). Studies investigating this phenomenon highlight the fact that stereotypes are not necessarily activated by an auditory stimulus, but can be activated by what a listener believes that they are about to hear. For example, in one study, a group of teachers listened to an audio clip of a child speaking in French. Some teachers were told that the speaker was Swiss-German while others were told that he was Serbian. Those who believed the student was Swiss-German rated his language proficiency differently than those who believed the student to be Serbian, despite the fact that there was no difference in the auditory stimulus (Berthele 2012 : 464-465). In another study, university students listened to an audio clip of a woman giving a short lecture. Those who were shown a photo of an Asian woman heard a stronger accent and demonstrated a less accurate comprehension of the lecture than those who were shown a photo of a Caucasian woman (Rubin, 1992 : 518-519).

8 These studies showed that stereotypes affect the comprehension and attitude of a listener in unconscious ways. While they postulated a mechanism, they did not investigate how it might transfer to grading by trained or untrained judges.

1.2. Matched Guise Test to Measure Unconscious Influences

9 Results of studies that target LS and RLS suggest that bias in speech perception can be unconscious and automatic, as the brain selectively encodes and processes incoming information (von Hippel, W., Sekaquaptewa, D., and Vargas, P. 1995 : 180-184). The selection process is influenced by individual beliefs but can also be strongly affected by

dominant stereotypes circulating in a society. Associations between a certain social group and its stereotypical attributes can become overlearned and automatically activated in response to seeing or hearing someone believed to belong to that group (von Hippel, Silver & Lynch, 2000 : passim). Because the filtering happens unconsciously, listeners are often not aware of their own subjectivities.

10 As far as language is concerned, the method often used to explore LS and RLS is referred to as a matched guise test and is based on the classic study by Lambert et al., cited previously (1960). This method is useful for measuring stereotypes because it activates unconscious or automatic responses without explicitly telling the participant what is being measured. In LS studies, participants are typically exposed to several stimuli that are either produced by the same speaker (speaker controlled) or produced by several different speakers who say the same content (verbal guise). However, participants do not know this ; they believe that the stimuli are recorded by different speakers. In RLS studies, participants are exposed to just one stimulus but are informed differently about the identity of the speaker (speaker and content controlled). While elements of ecological validity are sacrificed, these methods offer a way of measuring effects of unconscious or automatic stereotypes on perception. This is difficult to measure with more explicit methods, because participants may not be aware of their subjectivities in perception (Garrett, Coupland & Williams 2003 : passim).

1.3. Research Questions

11 This study uses a matched guise test to explore the effects of the ethnicity of three speakers on the detection of grammar errors and global proficiency judgments by untrained judges. This experiment was conducted in France, and thus the particular cultural and linguistic context should be taken into consideration when formulating the research questions. In France, some researchers have posited the existence of four stereotyped social categories, which could be referred to as “ethnicities”. These categories are present in the social representation but are of course not recognized or sanctioned by any official institutions. These four categories are “Arabic”, “white”, “black” and “Asian” (Simon and Clement, 2006 :2-3). Of course, the fact that we use such categories does not mean that we adopt an essentialist view on them. Rather, we consider that ethnicity is dynamic, socially constructed and that it varies across time, place, who perceives and who is perceived (Richeson & Sommers, 2016). The use of these categories in this research is designed to help measure the effect that they may have, particularly at an implicit or unconscious level, on the judgments of those in the society around them.

12 Although ethnicity is a constructed category that works within the social mind, it may pave the way for prejudices which lead in turn to discrimination against a certain group at individual or societal level. Evidence suggests that in France, those seen as “Arabic” encounter widespread discrimination. For approximately five decades, France has seen an influx of immigrants from Maghrebian countries[1], primarily Arabic speakers. While this is a source of cultural richness for the country, it is also a source of ideological and political tension. Multiple studies have shown that these immigrants and their children born in France have more difficulty finding employment (Simon, 2003 : passim ; Silberman, Alba & Fournier, 2007 : 22-24) or are more likely to be stopped and questioned by police (Jobard & Levy, 2010).

13 This particular cultural context informed our choice of speech stimuli as well as our decision to target the effects of the "Arabic" stereotype on speech perception. We posed the following two research questions :

1. When participants listen to one of three speech stimuli containing identical content but read by three different foreign speakers :

a) Is there a difference in the number of grammar errors detected ?

b) How is this measure associated with overall proficiency scores assigned ?

14 2. When participants listen to the same speech stimuli and are grouped by those who consider that the speaker was "Arabic" and those who do not :

a) Is there a difference in the number of grammar errors detected ?

b) How is this measure associated with overall proficiency scores assigned ?

2. Method

2.1. Overview

15 For this experiment, we recorded three foreign exchange students (Taiwanese, Brazilian, and Syrian) reading an identical passage in French containing intentional grammar errors. The students were all female, studying French at the University of Grenoble, a city in the French Alps. A total of 343 French university students, divided into three groups, listened to one of these clips per group. They wrote down errors they heard and assigned a global score for proficiency, pronunciation, and academic potential. The data were analyzed in two ways. The first compared differences in judgments across the three groups of participants, e.g., judgments of those who heard the Taiwanese speaker were compared to those who judged the Brazilian, and Syrian speaker. Since these judgments were influenced by real differences in the heard speech patterns, we referred to this as the analysis of the *objective* differences on judgments effect (Objective difference analysis). Differences between the three groups of judges were noted in mean number of grammar errors detected and in judgments of academic potential.

16 The second analysis used only the judgments of the participants who heard the Syrian speaker, comparing those who perceived her as "Arabic" and those who did not. Here, all judges heard the same recorded speaker, so judgments were only influenced by how they perceived the speaker. We referred to this as the analysis of the *perceived* difference (Perceived difference analysis). Differences between these two groups were noted in mean number of grammar errors detected and in judgments of overall proficiency and pronunciation (Figure 1).

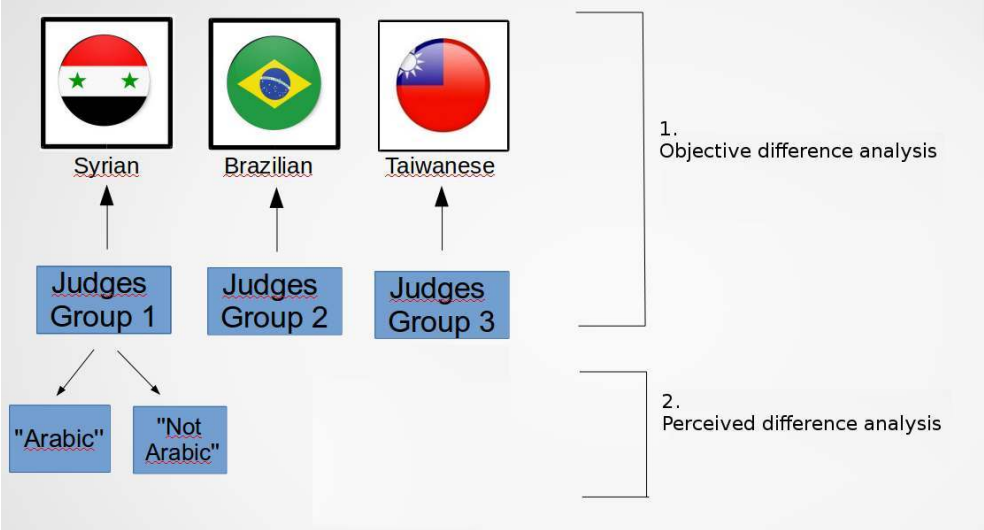

Figure 1 : Schematic representation of the objective and perceived difference analysis. A subset of the data used for analysis 1 were used for analysis 2.

2.2. Participants

17 Judges for this study were graduate and undergraduate students at the University of Grenoble in France. The age of the judges (N =343) had a mean (M) value of 20.2 years, with a standard deviation (SD) of 3.18. Data from judges were eliminated if they self-identified on a questionnaire that they either a) did not speak French as a first language, b) were not born in France, or c) fluently spoke the first language of the speaker they were listening to. The first two criteria were chosen to ensure that all judges were proficient in French language and grew up surrounded by French culture. The third criterion was chosen to eliminate the possibility that a judge who spoke the same language as the speaker they listened to might demonstrate some form of bias toward the speaker. After these eliminations, the final pool of valid judge data contained 313 samples.

2.3. Speech Stimuli

18 The stimuli for this experiment consisted of three audio clips, each of a different speaker reading an identical passage about a Parisian monument. The three speakers were similar or identical in age, gender, education, and intelligibility, chosen from eight possible clips. In order to ensure, as much as possible, that we chose three students with similar levels of intelligibility[2] when speaking in French, we recorded eight different speakers reading the same passage. We played these clips, in random orders, to 41 native French speakers, asking them to evaluate the intelligibility of the speakers on a 4-point scale (Appendix 1). No single evaluator heard all eight clips, but each clip was rated by at least 17 evaluators. Mean intelligibility scores were calculated and the three speakers with the closest scores were selected to generate the stimuli (Table 1).

19 After this selection, we re-recorded the three speakers but this time we inserted nine grammar errors into the passage (Appendix 2). To ensure that the errors were plausible, we analysed three French language learner corpora[3] and recorded and categorized

common types of errors. Nine errors found to be frequent in learners of varying linguistic backgrounds were chosen and three Masters students in French as a Foreign Language verified the plausibility of these errors in the context of the text in which they were inserted.

20 The speakers were recorded in an anechoic chamber. Care was taken to ensure consistency in intonation and speed. Final clips ranged from 52-55 seconds. These clips were verified by the same three FLE students to ensure that they did not contain any unintentional additional errors.

Table 1. Criteria for selection of three foreign students to record speech stimuli. Shaded rows indicate the selected speakers.

Speaker	Birthplace	First language	Gender	Age	Education	# of judges	Mean intelligibility score (1-4)
1	Switzerland	German	F	32	Masters student	41	3.27
2	Taiwan	Mandarin and Taiwanese	F	26	Masters student	41	2.34
3	Syria	Arabic	F	26	Masters student	41	2.25
4	China	Mandarin	F	29	Masters student	24	2.75
5	Germany	German	F	24	Masters student	24	3.17
6	Saudi Arabia	Arabic	F	23	Masters student	24	1.04
7	France*	Arabic	F	22	Masters student	41	3.98
8	Brazil	Portuguese	F	27	Masters student	17	2.24

2.4. Task Procedure

21 Judges were told that they were participating in a study to compare their ratings of a foreign student's language proficiency with ratings of trained judges. First, we played a 5-second audio clip to ensure that all participants could hear clearly. Next, we played one of the speech stimuli, once only. Judges were given ten minutes to complete the questionnaire, which prompted them to recall grammar errors and make overall judgments of proficiency, pronunciation, and academic potential.

22 Attached to the front of each questionnaire was a small 2x3 inch colour photo of the actual speaker. The photo was included to complement the audio stimuli with a visual image of the speaker. We felt that this format would be closer to a 'real life encounter' than with audio alone. Creating a more authentic interaction with video or in-person were beyond the scope of this study design. In order to standardize the images as much as possible, photos were head shots ; speakers had a neutral expression and clothing.

23 On the front of the questionnaire, judges were asked to write down a list of the grammar errors they detected in the clip. On the back, they were asked to rate the speaker on 4-point ordinal scales for : 1) overall language proficiency, 2) pronunciation, and 3) academic potential (Appendix 3). The scales for these three dimensions ranged from 1 : "strongly disagree" to 4 : "strongly agree".

24 Judges were not explicitly informed of the speaker's national identity or first language. To determine what a judge believed about the speaker's origins, they were asked to mark an "x" on a map to indicate where they believed the speaker to be from, writing the name of the country or zone below. We call this protocol 'Guess origin'. After completing their evaluation, the judges recorded biographical information including their date of birth, country of birth, age, first language(s), and additional languages spoken.

25 After completing about half of the data collection, a survey of responses revealed that almost no listeners identified the origins of the Syrian speaker as "Arabic"[4]. In order to ensure that we were able to complete the second analysis, measuring the effects of an "Arabic" stereotype, we changed the protocol for the remainder of the data collection. In our second protocol, which we call 'Explicitly informed', rather than giving participants a photo and asking them for the geographical origins of the speaker, the researcher told judges the first language of the speaker (Arabic, Mandarin, or Portuguese) before the clip began. Since only 3 of the 91 judges in the first part of the study guessed the speaker was Arabic, the rest were used as the "not Arabic" sample for the second analysis. This equalized numbers in the "Arabic" and "Not Arabic" conditions (Table 2).

Table 2. Number of participants in each dimension : Objective and Perceived difference analyses

Stimulus	**Syrian**		**Brazilian**		**Taiwanese**		
Protocol	Guess origin	Explicitly informed	Guess origin	Explicitly informed	Guess origin	Explicitly informed	Total
N	91	93	49	42	22	16	313
Perception	"Not Arabic"	"Arabic"					
N	84*	96**					180

***** of 91 who guessed origin, 7 were eliminated (3 guessed correctly and 4 did not complete the question)
** 93 were explicitly told, the 3 participants who guessed correctly were moved to this group

2.5. Analysis

26 Total number of errors detected in the speech stimuli followed a normal distribution. The mean number of errors were compared between speech stimuli (e.g., Syrian, Brazilian, Taiwanese) using a one-way analysis of variance (ANOVA), followed by a Tukey honest significant difference (HSD) post-hoc test. Pearson chi-square (x^2) tests were used to measure the association between speech stimuli and the identification of each error (yes vs. no) and each global score rating (ordinal scale) and the association between perceived dimension ("Arabic" vs. "not Arabic") and the identification of each error and each global score rating.

27 In all analyses, we hypothesized no difference between the speakers, with statistical significance at $\alpha = 0.05$. Data were analysed using SPSS for Windows (version 16.0, SPSS Inc, Chicago, IL).

3. Results

3.1. Guess vs Explicit Information

28 To determine whether the change in protocol between guessing and being told the origin of the speaker may have caused an effect, we compared numbers of errors detected and mean global scores across the two protocols for all listeners. There was no significant association between the mean number of errors detected and the mean global scores and the change of protocol for the Brazilian speaker.[5] In other words, judges rated her the same whether they knew she was Brazilian or not. However, there was a significant association in the case of the Syrian speaker. This suggested strongly that the change in protocol alone did not cause an effect ; rather, other factors were acting on the judgments.

3.2. Objective Difference Analysis (Syrian vs. Brazilian vs. Taiwanese)

29 This analysis examines our first research question : the differences in judgments across the three different speech stimuli. Each of the three speakers read the same passage containing identical errors, and judges heard one of the stimuli, noting specific grammar errors and overall impressions.

3.2.1. Grammar Errors

30 Overall, the mean number of grammar errors recorded by participants was 3.18, with a standard deviation of 1.30. All errors were detected at least once. There was a significant association between speech stimuli and number of errors detected, $F(2, 310) = 5.90$, $p = .003$, with participants detecting more errors in the Taiwanese clip than in the Brazilian clip ($p = .002$) or Syrian clip ($p = .010$) (Table 3).

31 The proportion of judges who detected each error differed by speech stimuli. Despite the fact that in general, more errors overall were detected in the Taiwanese student's speech, this phenomenon was inconsistent across each individual error (Table 4). In other words,

judges seemed to be better able to identify different errors in each of the three speech stimuli.

Table 3. Comparison of total errors detected across the three speech stimuli.

Speech stimuli	Mean	Median	Standard deviation
Taiwanese	3.82	4	1.136
Syrian	3.15	3	1.308
Brazilian	2.98	3	1.273

Table 4. Proportion of participants who detected each error across three speech stimuli. Shaded cells are stimuli that had the highest proportion of judges detect the error.

Error	Syrian	Brazilian	Taiwanese	df	Pearson chi-square	p
1	72.8	78.9	89.5	2	5.24	.073
2	43.5	24.2	15.8	2	16.74	.000
3	14.1	8.3	5.4	2	6.84	.033
4	15.2	34.1	15.8	2	13.80	.001
5	12.0	18.7	23.7	2	4.44	.108
6	5.4	8.8	60.5	2	86.07	.000
7	63.0	70.3	52.6	2	3.78	.151
8	38.0	12.1	52.6	2	26.97	.000
9	40.8	23.1	26.3	2	9.60	.008

3.2.2. Global Scores

32 Across all three ordinal-scale questions that measured proficiency, pronunciation, and academic potential, (Appendix 3), a greater proportion of participants strongly agreed that the Brazilian speaker had good overall proficiency, pronunciation, and potential. This difference was significant for potential only (Table 5). The order of the “strongly agreed” proportions – the Brazilian student highest, the Syrian student in between, and the Taiwanese student lowest – is the reverse of the mean number of errors detected in their speech (see Table 3).

Table 5. Proportion of judges who gave each score on the three, ordinal-scale questions

	4 : "Strongly Agree"	3	2	1 : "Strongly Disagree"	Pearson chi-square significance
"Is Proficient"					
Syrian	32.6	63.0	4.3	0.0	X^2 (4, N =313) = 7.30, p =.121
Taiwanese	26.3	73.7	0.0	0.0	
Brazilian	**44.0**	53.8	2.2	0.0	
"Has good pronunciation"					
Syrian	27.7	65.2	6.5	0.5	X^2 (2, N =313) = 7.322, p =.292
Taiwanese	15.8	81.6	2.6	0.0	
Brazilian	**31.9**	65.9	2.2	0.0	
"Has high potential"					
Syrian	30.4	64.7	3.8	1.1	X^2 (2, N =313) = 14.14, p =.028
Taiwanese	26.3	71.1	2.6	0.0	
Brazilian	**49.5**	45.1	5.5	0.0	

3.3. Perceived Difference Analysis (« Arabic » vs. « Not Arabic »)

33 This analysis addressed our second research question : all judges in this analysis listened to the exact same speech stimulus : that of the Syrian speaker. Approximately half of the judges perceived the speaker as "Arabic", and half did not. We compared the responses across these two groups to see if the category "Arabic" triggered any bias.

3.3.1. Grammar Errors

34 Of the participants who listened to the Syrian speech stimulus, those who were told or perceived the speaker to be "Arabic" detected, on average, fewer grammar errors (M = 2.76, SD = 1.25) than those who perceived her as "not Arabic" (M = 3.60, SD = 1.24), $t(178)$ = 4.47, p < .001. This pattern was consistent across each of the nine errors, but the association was only significant in four of the nine errors : 1, 4, 8, and 9 (**Table 6**). This effect reveals that when judges perceived the speaker as "Arabic", they were less likely to note grammar errors in the speech stimulus.

Table 6. Proportion of participants who found each intentional grammar error. Highlighted lines indicate statistically significant difference between judges who perceived the speakers "Arabic" and "not Arabic".

Error ID	Arabic group (% who found error)	Not Arabic group	Pearson Chi-square test (df = 1, N =180)	Exact sig. (2-sided)
1	64.6	**81.0**	5.98	.014
2	42.7	**46.4**	.25	.616
3	20.8	**27.4**	1.056	.304
4	8.3	**22.6**	7.17	.007
5	9.4	**14.3**	1.05	.306
6	4.2	**7.1**	.76	.384
7	61.5	**65.5**	.311	.577
8	30.2	**46.4**	5.01	.025
9	34.4	**48.8**	3.86	.050

3.3.2. Global Scores

35 The proportion of participants who strongly agreed that the speaker was proficient, had good pronunciation, and had academic potential, differed by perceived ethnicity. In all three questions, a greater proportion of participants in the group "not Arabic" judge the recorded speaker as more skilled and with more potential. This effect was significant for overall proficiency (p =.003) and pronunciation (p =.001) (Table 7). In other words, although judges in the "Arabic" group noted fewer quantitative grammar errors, they rated the speaker less positively in their global evaluations of her ability, whereas the opposite could have been expected.

Table 7. Proportion of Participants who gave each score on the three ordinal-scale questions

	4 "strongly agree"	3	2	1 "strongly disagree"	Pearson chi-square significance
Proficient in the language					
"Arabic"	22.9	69.8	7.3	0.0	$c^2(2, N=180) = 11.47, p = .003$
"not Arabic"	**44.0**	54.8	1.2	0.0	
Good pronunciation					

"Arabic"	17.7	69.8	11.5	1.0	$c^2(3, N=180) = 16.20, p = .001$
"not Arabic"	**39.3**	59.5	1.2	0.0	
High academic potential					
"Arabic"	25.0	67.7	5.2	2.1	$c^2(3, N=180) = 5.09, p = .165$
"not Arabic"	**36.9**	60.7	2.4	0.0	

3.4. Negative Ratings

36 Few judges chose "disagree" or "strongly disagree" on the ordinal-scale questions. In fact, these choices represented only 41 of the 939 total scale questions answered (4.4 %). 30 of the 41 negative responses were given by judges listening to the Syrian speech stimulus, and strikingly, of the 30, 26 were given by participants who perceived the speaker to be "Arabic" (Table 8).

TABLE 8. Number of negative responses "disagree" or "strongly disagree" on the ordinal-scale questions.

Question	"Arabic" (n =96)	"Not Arabic" (n =84)
Is proficient	7	1
Good pronunciation	12	1
Academic potential	7	2
Total number	26 (9.02 %)	4 (1.58 %)

4. Discussion

37 Scores on tests like the TOEFL have important consequences for test takers and for society, helping to determine who gets into a university program or who becomes a citizen. As a society, we have a vested interest in ensuring that these scores are accurate and reliable. Although inter-rater reliability is well researched and documented, the potential for collective biases shared across a group of raters is cause for concern.

38 The matched guise test, which was developed to explore LS and RLS, is indirect, targeting unconscious or automatic responses. In LS studies, participants are typically exposed to several stimuli that are either produced by the same speaker (speaker controlled) or produced by several different speakers who say the same content (verbal guise). In RLS studies, participants are exposed to just one stimulus but are informed differently about the identity of the speaker (speaker and content controlled). While elements of ecological validity are sacrificed, these methods offer a way of measuring effects of unconscious or

automatic stereotypes on perception. This is difficult to measure with more direct methods, because participants may not be aware of their subjectivities in perception (Garrett, Coupland & Williams 2003 : passim).

39 In this exploratory study, we examined two dimensions of evaluation bias : the evaluation bias resulting from objective differences in the heard utterances recorded by three foreign students from three countries (Syrian, Taiwanese, and Brazilian) and the evaluation bias resulting from the perception of the same recorded speaker and her assignation to two different ethnic origins ('Arabic' vs 'Not Arabic').

40 We found interesting results in both dimensions of this study. In the first dimension, differences in evaluations were noted across the three different speech stimuli. More grammar errors were identified in the Taiwanese student's speech. Addressing the first part of our research question, this finding confirms that there was a difference in grammar errors detected across the three speakers. It is difficult to determine, in the scope of this study, whether the difference was due to accent, intonation, listener expectations, or other differences between the three stimuli.

41 On the section of the questionnaire requiring global scoring, participants consistently rated the Taiwanese speaker less positively than the other two speakers on all three questions. The fact that this difference did not reach statistical significance in all questions was likely due to the unequal and non-parametric distribution of responses, resulting in small, irregular samples for the "disagree" and "strongly disagree" categories.

42 Our findings indicate that the number of grammar errors was correlated to overall proficiency scores, with judges finding the most grammar errors in the Taiwanese students' speech, and also rating her the lowest in overall proficiency ratings. This correlation is logical and suggests in this context, the more grammar errors perceived, the lower the judges' impression of overall language ability would be.

43 The results for the second dimension of this study, addressing the effect of a perceived "Arabic" identity on judges' scores, contradicts the trend seen in the first dimension.

44 Although all participants listened to the same audio clip of the Syrian speaker, those who guessed or were informed that she was "Arabic" tended to score her more negatively on subjective criteria like proficiency, despite detecting fewer grammar errors in her speech. Differences were often large, as with the fourth error, where almost three times as many participants detected the error in the "not Arabic" group (22.6 % vs 8.3 %).

45 This contradictory tendency to score the "Arabic" speaker *less* positively while detecting *fewer* actual errors in her speech is not logical, or consistent with our findings in the first part of this study. One would expect that fewer grammar errors would correspond with a more positive overall rating of language proficiency, as it did with the evaluations of the Taiwanese and the Brazilian student (judges found fewer errors in the Brazilian student's speech and consistently rated her more favourably than they did the Taiwanese student). One possible explanation for this is that the activation of a negative "Arabic" stereotype caused judges to allocate fewer cognitive resources to the listening task. The listening task was highly demanding, with speakers making nine errors in just under one minute, and the judges were only given one chance to listen. They were informed that the speaker was "Arabic" only a few seconds before the clip began, and it is possible that this knowledge activated thoughts or emotions that prevented them from as effectively hearing the errors. On the other hand, on the second page of the questionnaire, when

judges gave global ratings of mastery, pronunciation, and potential, they had more time to think and their responses more logically reflected a bias.

46 The result that most convincingly points to an "Arabic" bias is the distribution of negative ratings on the global evaluations (mastery, pronunciation, and potential). Although most of the scores were positive ("strongly agree" and "agree"), the few negative scores were very unequally distributed across listeners of the Syrian speaker. Of 30 negative responses to this speech stimulus, 26 were given by participants who believed the speaker was "Arabic". This discrepancy suggests that the perceived ethnicity of the speaker as "Arabic" did influence participants' perception of her speech, provoking a more negative response in global evaluation questions and possibly preventing them from effectively hearing the errors.

47 Taken as a whole, the results of this study suggest that ethnicity, whether real or perceived, can play a role in influencing quantitative and qualitative judgments of language proficiency. Differences were noted between objectively different speech stimuli as well as between groups of listeners who heard the same speech stimulus but perceived the ethnicity of the speaker differently. This strongly suggests that it is not only a speaker's delivery that influences the evaluation but also a listener's pre-conceived notions and beliefs about the speaker. Evaluation of language proficiency is a subjective, two-way process, influenced by both linguistic and non-linguistic factors. In this study, we discovered differences in speech perception at both the phonetic and word and sentence level.

5. Conclusion

48 As a society, we have a vested interest in ensuring that scores on high-stakes language proficiency tests are reliable. Measures of reliability must include inter-rater reliability but also consider the potential for "reliable" biases, shared across a majority of raters of a single candidate. The seemingly infinite number of variables that play a role in the evaluation process present challenges to these types of measurements. The matched guise test offers a way of beginning to explore these phenomena. While results from this study cannot be directly transferred to real-life situations such as the TOEFL oral interview, they highlight the need to more closely examine societal stereotypes and the role they might play in judges' ratings. This information could help to inform rater training and policy decisions about language testing.

Appendix 1

49 Choices for intelligibility scores (translated from French by the author)

1	Perfect intelligibility - as easy to understand as if I was listening to a native French speaker
2	Very good intelligibility - easy to understand
3	Somewhat good intelligibility - I can understand everything with a little effort
4	Poor intelligibility - somewhat difficult to understand

Appendix 2

50 Grammar errors inserted into the text.

51 Complete text read aloud :

52 Le Panthéon est l'un des monuments les plus célèbres de Paris. Il est situé sur la montagne Sainte-Geneviève **à le** 5e arrondissement.

53 **Cette** monument est très grand et imposant. Il **est** en tout 83 mètres de haut. Il a un dôme **majestueuse** et trois coupoles. À l'intérieur, il y a une crypte et aussi des tableaux et des fresques illustrant la vie de sainte Geneviève. Bien qu'il **est** exposé au soleil, il n'y a pas beaucoup de fenêtres et il **est** donc froid et sombre à l'intérieur.

En 1885, le gouvernement a décidé **à** rénover le Panthéon. A partir du moment où les travaux **a** été terminés, il est devenu un temple laïc destiné à honorer les français célèbres et à **souvenir** des événements marquants de l'histoire de France

	Error	Corrected	Type of error
1	À le 5e arrondissement	**Au** 5e arrondissement	Contraction of preposition
2	Cette monument	**Ce** monument	Gender of determiner
3	Il est en tout	Il **fait** en tout	Verb choice
4	Un dôme majestueuse	Un dôme **majestueux**	Adjective agreement
5	Bien qu'il est exposé	Bien qu'il **soit** exposé	Choice of verb
6	Il est donc froid	Il **fait** donc froid	Verb choice
7	Décidé à rénover	Décider **de** rénover	Preposition choice
8	Les travaux a été	Les travaux **ont** été	Subject-verb agreement
9	À souvenir des événements	À **se** souvenir des événements	Pronominal verb choice

Appendix 3

54 Three global questions on questionnaire (translated from French by the author)

55 1. This student demonstrates a level of language proficiency which allows her to express herself clearly

56 2. Despite her accent, this student has an intelligible pronunciation which does not cause difficulty in understanding her.

57 3. According to her performance on this clip, this student would be able to follow a university curriculum in France.

BIBLIOGRAPHIE

Berthele, Raphael. 2012. “The Influence of Code-Mixing and Speaker Information on Perception and Assessment of Foreign Language Proficiency : An Experimental Study”. *International Journal of Bilingualism* 16(4) : 453-466.

Bradac, James J., Aaron C. Cargile, and Judith P. Hallett. 2001. “Language Attitudes : Retrospect, Conspect, and Prospect. In *The New Handbook of Language and Social Psychology*, edited by W. Peter Robinson and Howard Giles, 137-158. Chichester, England : John Wiley.

Educational Testing Service. “TOEFL IBT Scoring Guides (Rubrics) for Speaking Responses”. Accessed September 2, 2014. http://www.ets.org/toefl/institutions/scores/guides

Garrett, Peter D., Nicolas J. R. Coupland and Angela M. Williams. 2003. *Investigating Language Attitudes : Social Meanings of Dialect, Ethnicity and Performance.* Cardiff, UK : University of Wales Press.

Hay, Jennifer and Katie Drager. 2010. “Stuffed Toys and Speech Perception”. *Linguistics* 48(4) : 865-892.

Jenkins, Susan and Isabel Parra. 2003. “Multiple layers of Meaning in an Oral Proficiency Test : The Complementary Roles of Nonverbal, Paralinguistic, and Verbal Behaviors in Assessment Decisions”. *The Modern Language Journal* 87(1) : 90-107.

Kang, Okim. 2008. “Ratings of L2 Oral Performance in English : Relative Impact of Rater Characteristics and Acoustic Measures of Accentedness”. PhD diss., University of Georgia.

Kang, Okim and Donald L. Rubin. 2009. “Reverse Linguistic Stereotyping : Measuring the Effect of Listener Expectations on Speech Evaluation. *Journal of Language and Social Psychology* 28(4) : 441-456.

Jobard, Fabien and René Lévy. 2010. “Les contrôles d’identité à Paris”. *Questions pénales*, CNRS-Ministère de la Justice 23 (1) : 1-4. <halshs-00550222>

Lambert, Wallace E., Robert C. Hodgson, Robert C. Gardner, and Samuel Fillenbaum. 1960. “Evaluational Reactions to Spoken Languages”. *Journal of Abnormal and Social Psychology* 66(1) : 44-51.

Lindemann, Stephanie and Nicholas Subtirelu. 2013. “Reliably Biased : The Role of Listener Expectation in the Perception of Second Language Speech. *Language Learning* 63(3) : 567-594.

Lippi-Green, Rosina. 2012. *English with an Accent : Language, Ideology, and Discrimination in the United States.* New York : Routledge.

Nguyen, Beatrice B.D. 1993. “Accent Discrimination and the Test of Spoken English : A Call for an Objective Assessment of the Comprehensibility of Nonnative Speakers”. *California Law Review* 81 (5) : 1325-1361.Richeson, Jennifer A. and Sommers, Samuel R. (2016). Toward a Social Psychology of Race and Race Relations for the Twenty-First Century. Annual Review of Psychology, 67, 439–463.

Rubin, Donald L. 1992. “Nonlanguage Factors Affecting Undergraduates’ Judgments of Nonnative English-Speaking Teaching Assistants”. *Research in Higher Education* 33(4) : 511-531.

Rubin, Donald. 2012. “The Power of Prejudice in Accent Perception : Reverse Linguistic Stereotyping and its Impact on Listener Judgments and Decisions. Paper presented at the 3rd Annual Pronunciation in Second Language Learning and Teaching Conference, Iowa State University, September 16-17.

Silberman, Roxanne, Richard Alba and Irene Fournier. 2007. “Segmented Assimilation in France ? Discrimination in the Labour market Against the Second Generation. *Ethnic and Racial Studies* 30 (1) : 1-27.

Simon, Patrick. 2003. “France and the Unknown Second Generation : Preliminary Results on Social Mobility”. *International Migration Review* 37 : 1091-1119.

Simon, Patrick and Martin Clément. 2006. “Comment Décrire la Diversité des Origines en France ? Une Enquête Exploratoire sur les Perceptions des Salariés et des Étudiants. *Population & Societies* 425 : 1-4.

Smith, Larry and Nelson, Cecil. 1985. “International intelligibility of English : Directions and resources. *World Englishes, 4 :* 333-342.

Von Hippel, William, Denise Sekaquaptewa, and Patrick Vargas. 1995. “On the Role of Encoding Processes in Stereotype Maintenance”. *Advances in Experimental Social Psychology* 27 : 177-254.

Von Hippel, William, Lisa A. Silver, and Molly E. Lynch. 2000. “Stereotyping Against your Will : The Role of Inhibitory Ability in Stereotyping and Prejudice among the Elderly”. *Personality and Social Psychology Bulletin* 26(5) : 523-532.

NOTES

1. Tunisia, Morocco, Algeria, Mauritania, and Libya
2. By intelligibility, we are referring to Smith and Nelson’s (1985) definition of the ability of a listener to recognize individual words or utterances in a speech sample
3. FFLOC (French Learner Language Oral Corpora), Double Je (French TV show with foreign contestants) and recordings of conversations between eight exchange students at the University of Grenoble.
4. Any country or region where Arabic is the dominant language spoken
5. Photo condition = (P) and explicitly informed = (I)
Global scales : Proficiency 1.6 (P),1.56 (I), p =0.711 ; Pronunciation 1.68 (P), 1.72 (I) p =0.7 ; Academic Potential 1.56 (P), 1.56(I) p =0.99
Mean number of errors detected : 2.76 (P), 3.24 (I), = 0.07

RÉSUMÉS

Vu les conséquences des examens sur les compétences en langue, on doit s’assurer de la fiabilité des épreuves par des accords inter-juge. Toutefois cette technique ne détecte pas des biais partagés par un groupe de juges, par exemple, des biais collectifs motivés par des stéréotypes ethniques. Trois étudiantes – syrienne, taïwanaise, brésilienne – ont été enregistrées en lisant un texte en français comportant neuf erreurs grammaticales. Les enregistrements ont été évalués

par 343 étudiants natifs du français, à qui on a demandé de relever les erreurs et de noter la compétence générale en français. Les juges qui pensent que la première locutrice est « arabe » la jugent différemment de ceux qui lui attribuent une autre origine. À partir d'une appartenance ethnique perçue, des stéréotypes et des représentations sociales seraient donc mobilisés et ils modifieraient l'appréciation de la maîtrise d'une langue étrangère par des juges non formés.

Given the important consequences of oral language proficiency tests, it is important to ensure the reliability of the scoring by inter-raters. However, this measure doesn't identify biases shared by a group of judges, for example, collective bias triggered by stereotypes about the ethnicity of the speaker. Three foreign students (Syrian, Taiwanese, and Brazilian) were recorded reading a text with nine grammatical errors. The recorded texts were assessed by 343 native French speaking university students who were asked to note the errors and overall proficiency in French language. Judges who thought that the first speaker was 'Arabic' judged her differently than those who attributed a different origin to her. Based on the perceived ethnic affiliation, stereotypes and social representation could thus be triggered and affect the judgments of foreign language proficiency by untrained judges.

INDEX

Mots-clés : évaluation des capacités langagières ; biais ; accent étranger ; apprenant ; stéréotype ; ethnicité ; langue française
Palabras claves : language proficiency assessment ; bias ; foreign accent ; learner ; stereotype ; ethnicity ; French language

AUTEURS

CLAIRE GILCHRIST

Laboratoire Lidilem, Université Grenoble Alpes, France

JEAN-PIERRE CHEVROT

Laboratoire Lidilem, Université Grenoble Alpes, France