Proceedings of the 32nd North American Conference on Chinese Linguistics (NACCL-32). 2020.
Edited by Kaidi Chen. University of Connecticut, Storrs, CT. Pages 521-538.

Attitudes toward L2 Mandarin Speakers of Chinese and Non-Chinese Ethnicity

Robert Squizzero
University of Washington

Perceived ethnicity is known to bias perception of accentedness of non-native speakers, but most existing research demonstrating this has been done on majority languages of North American and European countries. In the present study, a group of L1 Mandarin listeners were asked to rate the personalities and language abilities of highly proficient L2 Mandarin speakers of Chinese and non-Chinese ethnicity after listening to short statements by each speaker. Results show an effect of ethnic ingroup favoritism on several personality traits but no difference in perceived language proficiency. In the same way that ethnic biases against L2 speakers of European languages have been confirmed by research, this study confirms that ethnic biases can be present in an East Asian context. The study also provides evidence that achieving nativelike pronunciation in a second language does not eliminate discrimination against members of an outgroup.

0. Introduction

Studies of perceived comprehensibility, intelligibility and accentedness[1] of second language (L2) speakers are often done on English in the US, where multilingualism isn't the norm and ethnic diversity is comparatively common. However, in much of the world, such as in many Chinese-speaking communities, multilingualism is the norm, yet ethnic diversity is not. Mandarin is spoken as a second language in greater China by many first language (L1) speakers of other Chinese languages, so exposure to accented Mandarin is not unusual. However, mutually unintelligible Chinese languages are popularly considered to be dialects of the same language, at least in mainland China, which results in a distinction between L2 Mandarin speakers of L1 Chinese languages and those of other L1s. While there are probably fewer proficient Mandarin speakers of non-Chinese L1s, many Mandarin speakers encounter these speakers through media, such as in unscripted television programs or internet videos, and some encounter these speakers in daily life. This exposure is likely to increase in the future with the growth in interest in Mandarin language

[1] *Intelligibility, comprehensibility,* and *accentedness* are used here as defined by Munro, Derwing, & Morton (2006). *Intelligibility*: Actual understanding of the meaning of a word or utterance, i.e. more understanding on the part of the listener is equivalent to more intelligibility of a part of a word or utterance. *Comprehensibility*: Ease with which a word or utterance can be understood, i.e. less effort on the part of the listener is equivalent to more comprehensibility on of a word or utterance. *Accentedness*: The degree to which the pronunciation of an utterance sounds different from an expected production pattern.

programs. In the United States, Mandarin is the 4th most popular foreign language by enrollment in K-12 education (American Councils for International Education 2017) and 7th most popular in higher education (Looney & Lusin 2018).

L2 speakers face specific consequences in society, such as negative treatment in the media, courts, housing, and employment (Lippi-Green 2012). While comprehensi-bility of L2 speech is influenced by morphosyntactic and lexical features, the factor most associated with L2 accentedness is pronunciation (Saito, Trofimovich, & Isaacs 2016). Both L2 speakers and their listeners often believe that it is possible to eliminate an L2 accent and may even believe that it is laziness on the part of the L2 speaker that prevents elimination of their accent (Gluszek & Dovidio 2010). There are conflicting opinions as to whether it is possible for adult L2 language learners to achieve nativelike pronunci-ation, but there is evidence that comprehensibility is improved if pronunciation instru-ction is more explicitly included in language class-room activities (Derwing & Munro 2015, Yang 2016). There is also evidence, outlined below, that perceived characteristics of a speaker's identity influence processing of the language input received by listeners. Some scholars argue that L2 speakers should not be responsible for "improvement" of their pronunciation to improve communication between themselves and L1 (or other L2) speakers, and that listeners should make an effort to better comprehend accented speech (Gluszek & Dovidio 2010, Kang, Rubin, & Lindemann 2015). Indeed, there is evidence that practice in listening to multiple accented L2 speakers of a language can improve speaker-independent intelligibility of accented speech (Baese-Berk, Bradlow, & Wright 2013. Regardless, students of second languages can make better-informed decisions about how much time and effort to spend on pronunciation with accurate information on how their accented speech is perceived.

1. Literature Review

Perceived speaker ethnicity is known to influence processing of and opinions about speakers and their speech. In a well-known study (Rubin 1992), college-student subjects listened to identical speech samples presented by a university instructor, but some subjects were presented with a photograph of a Caucasian woman while others were presented with a photograph of an Asian woman. Participants shown the photograph of the Asian woman performed worse on an intelligibility test and reported hearing a stronger accent. In a verbal guise study, Yook & Lindemann (2013) demonstrated that knowing a speaker's ethnicity can impact opinions about a speaker's character by surveying Korean college students who listened to speakers of five varieties of English. One group of participants were informed of the speakers' nationalities and ethnicities while the other group was not. In comparison to uninformed listeners, informed listeners tended to rate European-American and Korean English speakers higher than the others on status/competence traits, and they rated British Australian and African American Verna-cular English speakers lower on social attractiveness traits than did uninformed listeners.

SOCIAL COGNITION AND DUAL PROCESSING. The Dual Processing theory of social cognition, which developed over many decades and is not credited to one particular individual (but see Fiske & Neuberg 1990 for one major Dual Processing model and Fiske & Taylor 2017 for a general overview), can be used to better understand evaluative reactions to language input. The dual modes of reacting to and processing stimuli are 1) automatic and 2) conscious, and the model allows for investigation of the interface between the two. This section discusses relevant psychology literature along with its application to language attitudes research.

Macrae & Bodenhausen's (2000) explanation of categorical thinking illustrates many of the details of automatic processing. The categorical thinking process is divided into the three potential steps of activation, application, and inhibition. Automatic category activation refers to the initial, subconscious process that occurs when first encountering a stimulus, like the auditory stimulus of L2-accented speech, or being told that the ethnicity of the person who produced that speech matches the ethnicity of the listener. Automatic category activation may or may not take place; for the purpose of efficiency, the encountered person is associated with a relevant social category if and only if the social meaning of the encountered person is perceived as relevant to current information processing concerns. For the example of speaker ethnicity and L2-accented speech, the speaker will be categorized according to the listener's beliefs about L2 speakers and/or people of the L2 speaker's ethnicity. This categorization then triggers other representations of the category as well as stored information about members of such a social group. Category application is the next step in the categorical thinking process. Once a recognized individual is categorized into a social group, as long as the information perceived is consistent with the stereotypes held by the perceiver, the stereotype category is applied (Fiske & Taylor 2017). This application affects the subsequent interaction between individuals in the same way that any other prime would (Macrae & Bodenhausen 2000).

Categorical thinking, however, is not always applied; certainly, a perceiver could process a speaker as an individual person instead of a representative of a social category. According to a great deal of work cited by Macrae & Bodenhausen (2000), categorical stereotypes are most likely used when a perceiver lacks motivation or is burdened with a shortage of time or a high cognitive load (e.g. due to increased effort required to process accented speech). In such a situation, categorical thinking allows for efficiency and less expended effort in the processing of input. The perceiver can shift focus and free up mental resources for parsing of potential unexpected information. If the automatic perce-ption, however, is inconsistent with the held stereotype, or if multiple social stereotypes are simultaneously activated, the perceiver may either attempt to classify the perceived in terms of a subtype that contains unique categories differentiating the perceived from the stereotype(s), or the perceiver may instead try to access exemplars of members of the stereotype category (Fiske & Taylor 2017). The last possibility, which could happen after either form of category application, is that the perceiver proceeds with conscious processing as executive function takes over (Macrae & Bodenhausen 2000).

According to Fiske & Taylor (2017), conscious processing can inhibit and control the associations made during the category activation above. Fiske & Taylor outline several potential motives for determining whether processing rises to the level of con-sciousness or not; two of these, self-enhancement and trusting ingroup, encourage use of automatic processes. People usually expect good things from their ingroup members, and negativity stands out, cognitively speaking, much more quickly than positivity. An L2 speaker may be more likely to be considered to be an outgroup member by an L1 speaker, as would a person of a different ethnicity. Macrae & Bodenhausen (2000) argue that people are not always successful at inhibiting stereotypical thoughts; a person may fail to do so out of a lack of awareness that the thoughts occurring in the mind are stereo-typical. If a person is, in fact, aware, they must employ a cognitively demanding process to identify and replace the stereotypic thoughts. Should a person be unable to cope with such demands, the person has now created a situation in which the stereotypical thoughts are highly accessible yet not inhibited, resulting in more stereotypic thoughts. Given that cognitive load is increased when processing accented speech (Derwing & Munro 2015), inhibitory failure in this situation may be expected.

THE POTENTIATED RECRUITMENT FRAMEWORK FOR IMPLICIT AND EXPLICIT ATTITUDES. Bassili & Brown's (2005) *potentiated recruitment* framework has been used successfully to model language attitudes in terms of input, processing and response (Preston 2017). In the context of the dual processing model, potentiated recruitment is appealing as it allows components of attitude creation and activation to be accessed both consciously and auto-matically. Bassili & Brown (2005) argue that "connectionist networks" of stored experi-ences, evaluations, attitudes and exemplars are a good fit for the parts that underlie impli-cit and explicit attitudes. Rosenberg's (1968) *attitudinal cognitorium* is used as a means of understanding the structure of these networks in relation to stimulus processing and attitudinal response. The *cognitorium* is described as a connectionist network of *microconcepts*, each of which represents information sensitive to the context at the time it was stored in memory. A microconcept, a node in this network, is retrieved from memory in a new context each time it is accessed. It is then associated with other microconcepts, meaning that each implicit or explicit attitude a person holds can change in context depending on which microconcepts have been activated. This is why context is key to understanding attitudes, according to Bassili & Brown. These authors go as far as to say that "features of the context are just as involved in the potentiation features of the attitude object" (Bassili & Brown 2005, p. 555-556). Bassili & Brown's model is uniquely effect-tive in accounting for attitude malleability, as no node is likely to be recalled in the exact same way as when it was originally stored. Potentiation, or level of activation, of micro-concepts is influenced by a perceiver's recent information processing experiences, curr-ently perceived information about the attitude object, activation between linked concepts in the cognitorium, and, for explicit attitudes, cognitive activity in working memory. The influence from currently perceived information "exert[s] powerful

potentiating influences on the attitudinal cognitorium" (Bassili & Brown 2005, p. 553) and appears to be easiest to manipulate by a researcher in an experiment in the form of a prime.

SOCIAL COGNITION IN THE CHINESE CONTEXT. To apply a social cognition framework to perception of second-language Mandarin speakers by contemporary Chinese, certain concepts must be modified, or at least qualified, to fit the Chinese context. The concept of membership in an East Asian (typically Chinese, Japanese, or Korean) "collectivist" culture can be misinterpreted as loyalty to or identification with one's ingroup from the point of view of a western, "individualistic" culture. Yuki (2003) found that for small social groups of students in the US and Japan, Americans identified more with and were more loyal to their ingroups than their Japanese counterparts, according to psychological measures of ingroup loyalty and identity. Yuki also found that while knowledge of role relations within a group, as measured by a scale of subjective sociometric knowledge, was correlated with ingroup loyalty and identity for both Japanese and American respondents, perceived ingroup homogeneity was only correlated with ingroup identity for Americans. Yuki's experiment was set up from the viewpoint of traditional Confucian notions of social harmony as resulting from fulfillment of one's particular role in the family and not from loyalty to or identification with one's country, community, or society. In fact, according to Liu, Li, & Yue (2010), the idea of asserting one's identity with a broader social group, at least in China's Confucian culture, did not present itself until the forced establishment of European treaty ports in China. Ingroup favoritism with respect to membership in a social group broader than the family first became relevant and widespread in China via the adoption of western nationalism, done as a defense mechan-ism against Europeans and their stratified society, and eventually as "benevolent paternal-ism" towards outgroup members perceived as inferior (Liu et al. 2010). Western-style nationalism in modern China, however, is not necessarily a direct reaction to Western nations, given, for example, the widespread knowledge of the atrocities committed by Japan in China in the early twentieth century. Ethnicity, then, in China, has not historic-ally been divorced from nationality in terms of ingroup/outgroup categorization, and can serve as a source of discrimination in contemporary Chinese society. It has yet to be demonstrated, though, that ethnicity can interact with language in such categorization and subsequent discrimination.

LANGUAGE STEREOTYPES. Existing research on language stereotypes predicts that listeners evaluate speakers along dimensions of *status*, *solidarity*, and *dynamism*. These three affective dimensions are related to the types of affective meaning in social psychology that can be applied to evaluation of any object: *potency*, *evaluation*, and *activity* (Osgood, May, & Miron 1975). According to Zahn & Hopper (1985) and Fiske, Taylor, Cuddy, & Xu (2002) (the latter for *status* and *solidarity* only), individuals rated highly on *status* are thought to be of high competence, high social class, and high language fluency, people rated highly on *solidarity* are thought to be warmer and more socially and aesthetically attractive, and those rated highly on *dynamism* are thought to have higher levels of activity and social power. While Osgood et al. (1975) found that there is some

cultural variation in dimensions of affective meaning, they also found that evaluation, potency, and activity were significant for Cantonese speaker-listeners. Wible & Hui (1985) and White & Li (1991) found that evaluation, potency, and activity were significant for L1 Mandarin speakers listening to L2 Mandarin speech.

The Matched Guise Technique (Lambert, Hodgson, Gardner, & Fillenbaum 1960) traditionally presents a group of listeners with several recordings of the same text, each ostensibly produced by a different speaker. In actuality, some of the recordings are sets of two with each set produced by the same speaker using a different language or language variety. After listening to each recording, respondents are asked to rate each speaker's voice on semantic differential scales of personality traits and language impressions that are expected to represent listeners' attitudes of the groups thought to use these different language varieties. Use of the same speaker removes the possibility of ratings based on idiosyncratic variation naturally found between speakers. In contrast to traditional matched guise studies, the language samples presented to respondents are not varied in the present study (respondents rated the exact same language samples twice); rather, different information about the speaker's ethnicity is presented to respondents for each "guise." A similar technique been used with speaker nationality in previous matched guise studies (Hay, Nolan, & Drager 2006; Niedzielski 1999).

2. Research Questions

This research study aims to answer the following questions: Do ethnically Chinese native Mandarin-speaking listeners find ethnic Chinese L2 speakers to be more socially attract-ive and/or less accented due to simple ingroup bias, i.e. ethnic solidarity? Conversely, do these listeners find highly proficient ethnic outgroup members to be more socially attract-ive and/or less accented due to listener ideologies that ethnic ingroup members should know how to speak the language? Or, is the effect of ethnicity more nuanced than this?

3. Methods

STIMULI. L1 Mandarin stimuli were used to test the hypothetical situation in which an L2 speaker has acquired a nativelike pronunciation of Mandarin. Stimuli were selected from the Mandarin Affective Speech Corpus (Yang, Li, & Wu 2007), which contains recordings of 68 students (males: 45, mean age = 21.7) at Zhejiang University, all of whom had lived in Mainland China since birth and the "majority [of which] were trained to speak in standard Mandarin from early childhood" (p. 3). This corpus was chosen because it contains a large number of speakers reading the same sentences and is control-led for emotional state. The present study used 20 "neutral" recorded utterances from 20 different speakers (males: 10), chosen by the author to be relatively consistent in terms of prosody (length of utterance) and recording quality (presence of clipping, signal-to-noise ratio). In addition to the author, two L1 speakers of Mandarin from two different regions of China and one other highly proficient L2 Mandarin speaker screened the stimuli for regionally

marked speech features (such as merger of alveolar and retroflex fricatives, vowel diphthongization, etc.). Four speakers were presented as matched guises, while the remaining sixteen utterances were included as distractors. The example speakers and sentences for the matched guises are shown in Table 1. Different target sentences were used to make the task more engaging and better maintain rater attention. Half of the recordings were (falsely) presented to respondents as having been produced by ethnically Chinese L2 speakers and the other half of the recordings were (falsely) presented to respondents as produced by ethnically non-Chinese L2 speakers, but there were no actual linguistic differences between the distractor sentences in each group presented to listen-ers, and there were no differences at all between the target (matched guise) sentences.

Speaker ID	Sex	Sentence - Chinese	Sentence – English Translation
F1	Female	今天晚上会下雨。	It's going to rain tonight.
M1	Male		
F2	Female	我们室友总是把寝室弄得很脏。	Our roommates always make the dorm room very messy.
M2	Male		

Table 1: Target sentences and speakers

RESPONDENTS. Respondents were required to have been raised in Mainland China, of self-reported native or near-native Mandarin proficiency, at least 18 years of age, and with no known speech or hearing disorders. Because the web survey was administered using Google Cloud Platform, which is often inaccessible in China, respondents were recruited in Seattle through emails disseminated to Chinese student groups and students in introductory linguistics classes, flyers posted on the University of Washington campus, and through researcher contacts, though trained sociolinguists and first-order researcher contacts were asked to refrain from taking the survey. The survey was distributed from 9/27/2019 – 11/06/2019. 30 of the 58 respondents evaluated all 24 audio stimuli, and 28 of the 30 respondents provided their demographic information. Analyses include the 30 respondents who evaluated all 24 audio stimuli, and the completion rate is calculated to be 52%. An additional four respondents completed enough of the survey to evaluate both guises of at least one speaker; these observations were also included in the analyses. Respondents ranged in from 18-36 years of age, with a mean of 22.9 years, a variance of 22.5, and a standard deviation of 4.7. 13 of the respondents were women, 15 were men, and none were of other genders. Reported ethnicities were Han (n=24), Manchu (n=2), Hui (n=1), and no response (n=1). 11 respondents reported being an L1 speaker of at least one language or dialect in addition to Mandarin. Respondents were raised in a mixture of hometowns from different geographic areas of China (see Online Supplemental Materials).

PRESENTATION. Respondents were told (in Mandarin) that researchers were interested in getting their impressions on how second language Chinese speech sounds. Prior to beginning the study, respondents were asked for informed consent. Respondents

were then told that they would be listening to 24 different speech samples, each recorded by a different, highly proficient, L2 Mandarin speaker. As to not give the impression that some stimuli were produced by heritage speakers, respondents were told that none of these speakers were raised in a Chinese-speaking country, none began studying Chinese before the age of 14, and none learned Chinese from their families. Respondents were told that half of the speakers were ethnically Chinese (华裔 *huáyì)*, and half were of non-Chinese ethnicity (非华裔 *fēi huáyì)*. Respondents were accurately told that they would randomly be assigned to first hear the group of twelve ethnic Chinese speakers (n=21) or the group of twelve non-ethnic Chinese speakers (n=13). After rating each block, respon-dents were advised to take a break of a few minutes before being presented with the remaining twelve speech samples. Respondents were told that they could listen to each sample as many times as they wanted but were asked to quickly respond to each question according to their first impression of each speaker. Respondents were asked to wear earphones for the study and to adjust the volume to a suitable level before beginning.

Respondents were told that they would be asked to answer several questions about each speech sample. In order to familiarize respondents with the task, they were provided with an example question containing 7 radio buttons on a scale of 1 to 7, with two example adjectives that would later be included in the actual rating task: 冷淡的 *lěngdàn de* "cold (personality)" written to the left of the scale and 热情的 *rèqíng de* "warm (personality)" to the right of the scale. They were told the colder their impression of a speaker, the closer to "cold" they should click, whereas the warmer their impression of a speaker, the closer to "warm" they should click. They were told that if they were unsure or had no impression of the speaker that they should select the radio button in the center, labeled as 没什么印象 *méi shénme yìnxiàng* "no impression." At the top of the next page, the speaker number and ethnicity of the speakers in the current block were displayed. For the first recording, respondents clicked a "play" button to listen to each recording. Below the media player, they were presented with twelve semantic differential scales corresponding to nine personality traits and three language measures, described in the following section of this paper. Respondents rated the speaker on each trait and then clicked the "next" button, which submitted their response, cleared the form, and played the next recording. After listening to and rating all stimuli, respondents were asked for demographic information and thanked for their participation.

Participants rated each guise on nine personality traits and three impressions of language proficiency along seven-step semantic differential scales, in keeping with Osgood et al. (1975). The speaker traits included on the questionnaire were chosen after consulting language attitudes work for L2 speakers of English (Yook & Lindemann 2013) and L1 speakers of Mandarin (Liao 2008, Lin 2018, Peng 2016, Tan 2016, Yang 2014). Raters were also asked to rate each speaker's standardness, fluency, and foreign-accentedness on seven-level scales. Four rating scales of speaker traits and language proficiency were selected for each evaluative dimension to which they were predicted to correspond: *status:* competent-incompetent, smart-stupid, standard-not standard, and fluent-not fluent,

solidarity: deep-shallow, warm-cold, likeable-unlikeable, and accented-not accented, and *dynamism:* strong-weak, polite-rude, modern-conservative, and kind-arrogant. All seven values of the semantic differential scales were used by one respondent in a pilot study, so the seven options were left in. All semantic differential scales were arranged by placing the trait with negative valence on the left and positive valence on the right, except for the accented/not accented scale, which was reversed in order to detect straightlining, a strategy in which respondents simply select the same answer to all questions in order to finish the survey faster. No straightlining was observed, and accented/not accented scores were transformed to match the other scales, i.e. a score of 1 was converted to 7, a score of 2 was converted to 6, and a score of 3 was converted to 5.

4. Results

Table 2 shows summary statistics in the form of medians and interquartile ranges (IQRs) by semantic differential scale, pooling ratings from all respondents on all target speakers (distractor recordings were excluded). These descriptive statistics establish that ratings differed in terms of perceived speaker personality and perceived speaker language pro-ficiency, as found by Wible & Hui (1985). While the median ratings for personality trait scales were either 5 (for arrogant/kind, rude/polite and unlikeable/likeable) or 4 (for the other seven personality traits), the median ratings for each of the language proficiency scales (accented/not accented, fluent/not fluent, standard/not standard) was 7. The IQRs for conservative/modern and weak/strong show that a very high proportion of respond-ents were unwilling or unable to evaluate the speakers on those scales. Visual distribu-tions of ratings are presented in Figure 1. An exploratory factor analysis was conducted via principal components analysis with varimax rotation and Kaiser normalization. As expected, this analysis yielded three factors with eigenvalues above 1, accounting for 65.89% of the variance. The factor loadings are shown in Table 3. Rotated component 2 clearly consists of the three scales on which language proficiency was rated. As has been observed in previous studies on social stereotyping, it may be the case that *dynamism* is not a dimension at play here. If so, it is likely that the scales intended to measure dynam-ism actually measured *status* (strong-weak) and *solidarity* (the remaining three), and that rotated component 1 represents *solidarity* and rotated component 3 represents *status*. The principal components analysis and the vastly different distributions in ratings indicate that it would be illogical to group the twelve scales into the original three indices of *status, solidarity* and *dynamism*, in which ratings of language proficiency and personality traits were combined. Further, the non-normal distribution of the ratings as shown in Figure 1 indicates that non-parametric statistical procedures should be used, and that combining scales into indices representing dimensions of affective meaning may obscure individual differences among respondents in their interpretation of the points on each semantic differential scale. The ratings in this study, then, are conceptualized as follows: each respondent k generates score x_{ijk} by rating speaker i on bipolar scale j. Raters evaluated

the same speaker on the same bipolar scales twice, one in each ethnicity condition, so their scores were compared as:

$$H_0: Chinese_{ijk} = Non\text{-}Chinese_{ijk}.$$

Scale (English)	Scale (Mandarin)	Median	IQR (25%-75%)
Not Accented / Accented	没/有洋腔洋调	7*	6-7*
Arrogant / Kind	高傲的/和爱的	5	4-5
Cold / Warm	冷淡的/热情的	4	4-5
Conservative / Modern	保守的/前卫的	4	4-4
Not Fluent / Fluent	不/流利的	7	6-7
Incompetent / Competent	没/有才干的	4	4-5
Rude / Polite	没/有礼貌的	5	4-6
Shallow / Deep	肤浅的/有深度的	4	4-5
Not Standard / Standard	不/标准	7	6-7
Stupid / Smart	愚蠢的/聪明的	4	4-5
Unlikeable / Likeable	不/招人喜欢的	5	4-6
Weak / Strong	不/坚强的	4	4-4

Table 2: Medians and interquartile ranges for target stimuli ratings, across raters and speakers.* indicates a transformed rating

Scale / Component	Predicted dimension of affective meaning	Rotated Comp 1	Rotated Comp 3	Rotated Comp 2
Rude / Polite	Dynamism	.73	.32	
Stupid / Smart	Status	.47	.63	
Shallow / Deep	Solidarity		.87	
Cold / Warm	Solidarity	.79		
Incompetent / Competent	Status		.78	
Weak / Strong	Dynamism		.69	
Conservative / Modern	Dynamism	.53		
Arrogant / Kind	Dynamism	.88		
Unlikeable / Likeable	Solidarity	.77	.34	
Not Standard / Standard	Status			.86
Not Fluent / Fluent	Status			.88
Accented / Not Accented	Solidarity			.74

Table 3: Principal component analysis with varimax rotation and Kaiser normalization

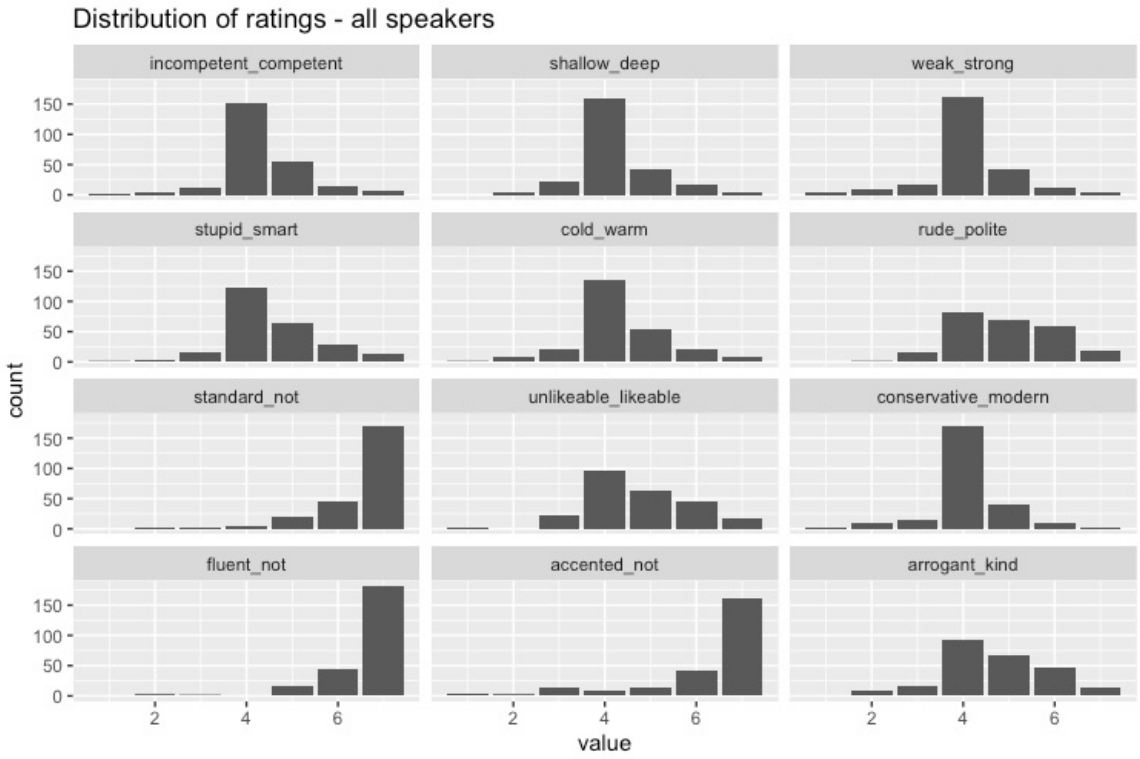

Figure 1 Counts of scores for target stimuli by rating scale, across raters and speakers. The columns, from right to left, represent status, solidarity, and dynamism measures.

Scale / Speaker	**F1**	**M1**	**F2**	**M2**
Not Accented / Accented	n.s.	n.s.	n.s.	n.s.
Arrogant / Kind	0.0094**	n.s.	n.s.	n.s.
Cold / Warm	0.0178*	n.s.	n.s.	n.s.
Conservative / Modern	n.s.	n.s.	n.s.	n.s.
Not Fluent / Fluent	n.s.	n.s.	n.s.	n.s.
Incompetent / Competent	0.0397*	n.s.	n.s.	n.s.
Rude / Polite	n.s.	n.s.	n.s.	n.s.
Shallow / Deep	0.0057**	n.s.	n.s.	n.s.
Not Standard / Standard	n.s.	n.s.	n.s.	n.s.
Stupid / Smart	0.0210**	n.s.	n.s.	n.s.
Unlikeable / Likeable	0.0013**	n.s.	n.s.	n.s.
Weak / Strong	n.s.	n.s.	n.s.	n.s.

Table 4: *p* values for each scale, by speaker. *$p < .05$, **$p < .01$

Scale	Pseudo-median	95% CI	Scale	Pseudo-median	95% CI
Arrogant / Kind	1.00	0.5-1.5	Shallow / Deep	1.00	0.00-2.00
Cold / Warm	1.00	0.00-2.00	Stupid / Smart	1.00	0.00-2.00
Incompetent / Competent	1.00	0.00-2.00	Unlikeable / Likeable	1.50	0.50-2.00

Table 5: Estimated location for the median of the difference from a sample of Chinese v. non-Chinese guises (pseudomedian) for speaker F1, all significant scales

To test whether these paired ratings were significantly different, a two-tailed Wilcoxon signed rank test was conducted, as in Bender (2004). This test has the advantage of not assuming that ratings on the semantic differential scales are continuous (interval or ratio) nor that the ratings are distributed normally in any of dimensions *i, j,* or *k*. In other words, it compares individual differences across conditions, so there is no potential for bias introduced by individual differences between raters' uses of the different scales on the different speakers. Results are shown in Table 4.

The *p* value of each two-tailed Wilcoxon signed rank test does not indicate directionality; it only indicates that there is a significant difference between rankings of each guise. To calculate directionality, an estimated location for the median of the difference between a sample from each guise is calculated based on a confidence interval, represented in Table 5. The sign of the estimated location indicates the direction in which rankings are higher; a positive value indicates that rankings in the first group (Chinese) are higher whereas a negative value indicates that the rankings in the second group (non-Chinese) are higher. All estimated location values for significant scale/speaker combinations were positive, indicating that the Chinese guise was rated more favorably for all statistically significant differences.[2]

3. Discussion

This study has provided strong evidence that L1 Mandarin-speaking listeners find ethnic Chinese L2 Mandarin speakers to be of higher status and more socially attractive than non-Chinese L2 Mandarin speakers due to ethnic ingroup implicit bias. In contrast to the results for character traits, no strong evidence is found for an effect of ethnicity on fluency, accentedness, or standardness. All observed differences in ratings for personality traits, statistically significant or not, are in the direction of favoring those who are ethnically Chinese. These findings confirm that ethnic biases present against L2 speakers of European languages can be extended to an East Asian context, despite claims by Liu et al. (2010) that such favoritism is only displayed by the Chinese as a defensive reaction to an outside threat or as a benevolent paternalism. While neither the presence of an outside

[2] Wilcoxon signed rank tests were also run to investigate ordering effects; no significant ordering effects were found. For more information, see the online supplemental materials.

threat nor the supposed inferiority of outgroup members was directly manipulated in the present experiment, no evidence for either is found in respondents' qualitative comments (see supplementary materials), and the semantic differential scales did not converge in a way comparable to paternalistic prejudice as previously demonstrated (Fiske et al. 2002).

The results, while robust, were only observed for the first of the four target speakers in this experiment that respondents heard. While interactions between demo-graphic variables such as speaker sex and ethnicity, or interactions between sentence length and demographic variables were suspected, it was not the case that an effect was shown for speakers of one sex or reading one sentence – it was shown only for a single speaker. One possible interpretation of this is that it reflects that social biases are not consistently applied; Fiske & Taylor (2017) write that categorical thinking does not happen consistently; rather, categorical thinking tends to occur in times of high cognitive load or low motivation (Macrae & Bodenhausen 2000). Higher cognitive load at the begi-nning of the survey due to unfamiliarity with the survey task is one possible explanation for why speaker F1 was the only speaker that was rated differently across guises.

The difference in distributions of personality and language proficiency ratings is also of interest to a theory of language attitudes; there is a high proportion of "no im-pression" responses found in all nine of the personality trait scales but a high proportion of "not accented," "not fluent," and "not standard" scores for language proficiency traits. Particularly high rates of half or more of the judgements being "no impression/don't know" were found for six of the nine personality trait scales: conservative/modern (169 of 246 ratings), weak/strong (161), shallow/deep (158), incompetent/competent (152), cold/warm (135), and stupid/smart (122). While the most common rating was "no impression/don't know" for the remaining three personality trait judgments, the median rating for rude/polite and unlikeable/likeable were "somewhat polite" and "somewhat likeable," whereas the median rating was still "no impression/don't know" for the arro-gant/kind scale. The most common rating and median rating on the language proficiency scales, in contrast, were the maximum ratings of "very fluent," "very standard," and "very unaccented." This means that respondents were not hesitant in judging a speaker's ability in the same way that they were hesitant about judging character. There are several possible interpretations for this difference. One interpretation is that social desirability bias prevents categorical thinking from applying to ratings of ability, i.e. it may be con-sidered fair, legally or morally, to judge language proficiency, as opposed to character. Alternatively, since most of the respondents are L2 learners of English who live in an English-speaking cultural context, they may be demonstrating empathy towards L2 learn-ers of Chinese. Another possible interpretation, supported by the qualitative comments, is that this population believes that language proficiency, unlike personality, can be judged by hearing a single sentence, such as those respondents heard in this study; this in itself is potentially relevant for future studies on ratings of L2 speakers. This is the opposite result, though, of what Wible & Hui (1985) found when L2 speakers of six different pro-ficiency levels were judged, which is that linguistic areas were less salient than person-ality traits. It may be the case, then, that

language proficiency is seen as something that can be controlled (Gluszek & Dovidio 2010), unlike personality traits, which may be perceived as innate. This would indicate that any preferential treatment of ingroup members is based on biased perceptions of character rather than biased perceptions of ability, at least for highly proficient second-language speakers. The fact that language proficiency ratings did not differ by ethnicity is of interest; while an effect was present for ethnic ingroup bias in rating speaker characteristics, no such effect was present for ratings of language proficiency for these highly proficient speakers in this study. This is particularly relevant for second language learners of Mandarin; it should be known that while speaking with a nativelike accent may result in being perceived as very fluent, standard, and unaccented, regardless of ethnicity, there may be still be negative treatment if a speaker is a member of the ethnic outgroup. Researchers investigating attitudes towards L2 speakers of any language should therefore be cautious about directly comparing ratings of language proficiency to those of personality. Chinese language teachers should also keep in mind that stereotypes against speakers may influence their proficiency. An L2 speaker may worry about how an outgroup member may stereotype them or may believe that an outgroup member may feign inability to understand the speaker. This stress may cause them to avoid communicating in the accented language, leading to less practice and less fluency, and arouse feelings of frustration or insecurity (Gluszek & Dovidio 2010).

The present study's methodology demonstrates a stricter adherence to the assumptions made by statistical procedures as compared to most existing research on language and social psychology. Responses along the semantic differential scales were treated as ordinal, resulting in reporting of median ratings and use of the Wilcoxon signed rank test. Treatment of data as ordinal eliminates the possibility of calculating mean ratings of speakers along the dimensions of affective meaning probed in this study, as well as the use of certain inferential statistical techniques, such as linear regression and ANOVA. Some researchers have argued that it is appropriate under certain circumstances to treat semantic differential scales as interval data (Carfino & Perla 2007); they are often considered to be special, separate from Likert scales or other opinion continua not validated for cross-cultural meaning. This view is contested, however, due to the potential for individual differences in use of the scales; such variation can be conceptualized as a three-dimensional space: variation between respondents in interpretation of the different anchor points on a scale, variation in interpretations of the distance between the seven points of different seven-point scales (scale variation), and variation in applying different scales to different stimuli (stimulus variation) (Murakami & Kroonenberg 2003). This individual variation can have the effect of obscuring, exaggerating, or overgeneralizing results when left unanalyzed. The present study's results should be interpreted with the understanding that individual differences in the ratings of different objects are often left uninvestigated. It may be the case that the result of the present study, in which only one of the four matched guise speakers is rated differently across guises, is not at all unusual; rather, it may be a result that is commonly obscured in other research by the pooling of ratings of different

stimulus speakers. More research in statistics and psychometrics would clarify best practices in quantitative analysis for research in language and social psychology; if semantic differential data can be safely treated as interval, it simplifies the analysis and increases inferential power. Nevertheless, if those who argue that it is only safe to treat these kinds of data as ordinal are correct, social psychologists and sociolinguists would do well to employ available statistical techniques for ordinal data, at least for within-subject study designs. The Wilcoxon signed rank test is an example of a powerful statistical procedure that allows for better understanding of individual distances while offering generalizable results.

While efforts were made to limit sampling bias, it is possible that some bias was introduced by the method by which respondents were recruited. The population sampled is one that lives in a multi-ethnic society, which is not the case for most Mandarin speakers, and is a young and educated population. Bias may also have been introduced by the dropout rate of 48%; reducing the length of the survey and using the semantic differential scales with fewer "no impression" ratings could assist in reducing this bias (Dillman, Smyth, & Christian 2014).

In addition to the short length of the utterances presented, it is possible that giving no information about each speaker other than ethnicity to respondents is too abstract to relate to a real-world situation and may not actually activate social categories or exemplars similar to the speaker being judged. Indeed, it has been shown that perceptions of L2 speech can vary even down to personae within a group (D'Onofrio 2019, Zhang 2008). It may be ideal to provide speaker demographic information in addition to ethnicity, even if kept constant across guises, such as the nationality of the respondents. Giving information about nationality was avoided, though, in order to avoid possible interactions between ethnicity and nationality. Another option, as suggested by one of the respondent's comments, is to tell respondents the context in which the example sentences were spoken; for instance, for the stimulus sentence about the weather, was this sentence produced in casual conversation, or was it produced by a meteorologist?

Future work measuring implicit, rather than explicit, attitudes or differences in processing of L2 Mandarin speech from speakers of different ethnicities may also help get to the root of the bias the population of L1 speakers from China. Such work would help determine whether social desirability bias is present, for example, by looking at differences in processing time when respondents believed they were listening to speech by ethnic Chinese and non-Chinese speakers.

4. Conclusion

The present work has demonstrated the potential for Chinese raters to express ingroup favoritism based only on knowledge of an outgroup member's ethnicity. For this population of raters, ingroup favoritism does not apply to an L2 speaker's language ability; rather, it applies to the speaker's personality. Chinese language students and teachers should be aware of the existence of this ingroup favoritism when considering how much

time and energy to devote to pronunciation. L2 Mandarin speakers of any ethnicity should know that positive or negative bias may very well occur based on their outward appearance or their interlocutors' understanding of their background, even if nativelike pronunciation is achieved, just as in the Western societies in which much research at the intersection of linguistics and social psychology is carried out. This is not to say that ethnic non-Chinese never experience preferential treatment in Chinese society; however, motivations for this treatment are probably not feelings of envy or admiration towards an outgroup (Fiske et al. 2002), nor are they an illustration of the Chinese concept of *chóng yáng mèi wài* (崇洋媚外), which refers to blind worship of foreign things. A detailed discussion of the source of the ethnic ingroup bias presently observed is beyond the scope of this investigation, though the nationalist ideology promoted by contemporary Chinese state media, as part of the global resurgence of nationalism also observed in the United States, Europe, and the Middle East, would not be expected to curtail ingroup favoritism nor outgroup derogation. Whatever the source of this bias, social scientists aiming to produce research which helps improve intergroup relations would benefit from a greater focus on the behavior of populations outside of the Western societies which are overrepresented in our disciplines.

SUPPLEMENTAL MATERIALS. Supplemental material is available at https://zeos.ling.washington.edu/publications/supplemental/squizzero/NACCLsupplemental.html. This material includes respondent demographic information, block-order analysis, and qualitative comments.

REFERENCES

American Councils for International Education. 2017. The National K-12 Foreign Language Enrollment Survey Report (June). Online: https://www.americancouncils.org/sites/default/files/FLE-report-June17.pdf

Baese-Berk, Melissa M., Ann R. Bradlow, and Beverly A. Wright. 2013. Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, *133*(3), EL174–EL180 https://doi.org/10.1121/1.4789864

Bassili, John N., and Rick D. Brown. 2005. Implicit and Explicit Attitudes: Research, Challenges and Theory. *The Handbook of Attitudes*, ed. by Dolores Albarracin, Blair T. Johnson, and Mark P. Zanna, 543–574. Mahwah: Lawrence Erlbaum.

Bender, Emily M. 2004. On the boundaries of linguistic competence: Matched-guise experiments as evidence of knowledge of grammar. *Lingua*, *115*(11 SPEC. ISS.), 1579–1598. https://doi.org/10.1016/j.lingua.2004.07.005

Carfino, J., & Perla, R. J. (2007). Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. *Journal of Social Sciences*, *3*(3), 106–116.

D'Onofrio, Annette. 2019. Complicating categories: Personae mediate racialized expectations of non-native speech. *Journal of Sociolinguistics*, 1–21.

Derwing, Tracey, and Murray Munro. 2015. *Pronunciation Fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam: John Benjamins.

Dillman, Don A.; Smyth, Jolene D.; and Christian, Leah Melani. 2014. Reducing People's Reluctance to Respond to Surveys. In *Internet, phone, mail, and mixed-mode survey: The tailored design method.* (pp. 19–55). John Wiley & Sons.

Niedzielski, Nancy A. 1999. The Effect of Social Information on the Phonetic Perception of Sociolinguistic Variables. *Journal of Language and Social Psychology*, *18*(1), 62–85.

Osgood, Charles E.; William H. May; and Murray S. Miron. 1975. *Cross-Cultural Universals of Affective Meaning*. Urbana, IL: University of Illinois Press.

Peng, Chun-Yi. 2016. *Talk Shows and Language Attitudes: A Sociolingusitic Investigation of Language Attitudes towards Taiwan Mandarin among Chinese Mainlanders*. The City University of New York.

Preston, Dennis. R. 2017. The cognitive foundations of language regard. *Poznan Studies in Contemporary Linguistics*, 53(1), 17–42. https://doi.org/10.1515/psicl-2017-0002

Rosenberg, Milton J. 1968. Hedonism, inauthenticity, and other goads toward expansion of a consistency theory. *Theories of cognitive consistency: A sourcebook*, ed. by Robert P. Abelson; Elliot Aronson; William J. McGuire; Theodore M. Newcomb; Milton J. Rosenberg; & Percy H. Tannenbaum, 73-111. Chicago: Rand McNally.

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. L., and Kim A. Smith. 1990. Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *International Journal of Intercultural Relations, 14*(3), 337–353.

Saito, Kazuya; Pavel Trofimovich; and Talia Isaacs. 2016. Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholingusitics*, *37*, 217–240.

Tan, Lekun. 2016. A Comparison of Language Awareness and Attitudes between Students in the Department of Taiwanese Literature and Other Departments. *Journal of Taiwanese Languages and Literature*, *11*(2).

Wible, David S, and C. Harry Hui. 1985. Perceived Language Proficiency and Person Perception. *Journal of Cross-Cultural Psychology*, *16*(2), 206–222.

White, Michael J., and Yun Li. 1991. Second-Language Fluency and Person Perception in China and the United States. *Journal of Language and Social Psychology*, *10*(2), 99–113.

Yook, Cheongmin, & Stephanie Lindemann. 2013. The role of speaker identification in Korean university students' attitudes towards five varieties of English. *Journal of Multilingual and Multicultural Development*, *34*(3), 279–296.

Yang, Chunsheng. 2014. Language attitudes toward Northeastern Mandarin and Putonghua (PTH) by young professionals. *Language and Discourse: An International and Interdisciplinary Journal*, *5*(2), 211–230.

Yang, Chunsheng. 2016. *The acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice*. Amsterdam: John Benjamins.

Yang, Yingchun; Li Dongdong; & Wu Zhaohui. 2007. Mandarin Affective Speech Corpus. Philadelphia: Linguistic Data Consortium. Online: https://catalog.ldc.upenn.edu/LDC2007S09.

Yuki, Masaki. 2003. Intergroup comparison versus intragroup relationships: A cross-cultural examination of social identity theory in North American and East Asian cultural contexts. *Social Psychology Quarterly*, *66*(2), 166–183.

Zahn, Christopher, & Robert Hopper. 1985. Measuring language attitudes: the speech evaluation instrument. *Journal of Language and Social Psychology*, *4*(2), 113–123.

Zhang, Qing. 2008. Rhotacization and the 'Beijing Smooth Operator': *Journal of Sociolinguistics*, *12*(2), 201–222. https://doi.org/10.1386/jmpr.7.1.43/1