

ORIGINAL ARTICLE

Searching for the “native” speaker: A preregistered conceptual replication and extension of Reid, Trofimovich, and O’Brien (2019)

Bianca Brown , Botagoz Tuzmagambet , Valentino Rahming , Chun-Ying Tu ,
Michael B. DeSalvo  and Seth Wiener* 

Department of Modern Languages, Carnegie Mellon University, Pittsburgh, PA, USA

*Corresponding author. Email: sethw1@cmu.edu

(Received 16 May 2022; revised 27 January 2023; accepted 5 February 2023; first published online 16 March 2023)

Abstract

This study conceptually replicated and extended Reid, Trofimovich, and O’Brien (2019), who found that native English speakers could be biased positively (or negatively) relative to a control condition in terms of how they rate non-native English speech. Our internet-based study failed to replicate Reid et al. across a wider population sample of “native” speakers ($n = 189$). Listeners did not change how they rated non-native English speech after social bias orientations and performed similarly across all five measures of speech and across age and race (Asian, Black, and Caucasian). We attribute our results to differences in the methods (in-person vs. online) and/or participants. Of note, roughly one-third of our “native” participants indicated proficiency in languages other than English and residency in 12 different English-speaking countries, despite identifying as a) fluent English speakers who b) used English primarily and c) acquired English before any other language from birth. These screening items taken together qualified “native” participants in line with traditional psycholinguistics research. We conclude that the concept of “nativeness” is tied to culture-specific perspectives surrounding language use. As such, the native/non-native categorical variable simultaneously serves and limits the advancement of psycholinguistics research.

Keywords: speech production; speech perception; accent; second language acquisition; social biases

Given that humans learn language at different ages, and to different degrees of proficiency, there will always be differences in how people speak a language. These differences result in an in-group/out-group dichotomy: native speakers (NS) and non-native speakers (NNS). Psycholinguistics research typically—though not always—considers a NS someone who learned the language as an infant (i.e., their first language) and to a high degree of proficiency. In contrast, a NNS learned

the language at a later age after another language was already acquired and often to a lower degree of proficiency than a NS.

Whereas this binary NS/NNS variable has undoubtedly pushed psycholinguistics research forward (e.g., Baese-Berk *et al.*, 2020; Baese-Berk & Morrill, 2019; Cristia *et al.*, 2012; Xie & Myers, 2017), speech perception research, like the speech signal itself, is inherently variable. The NS/NNS dichotomy often omits important details and diversity that researchers need to consider—or at the very least, report (Cheng *et al.*, 2021; Tsehaye *et al.*, 2021). These include the assumption of a “normal” process of first language acquisition (e.g., Abrahamsson & Hyltenstam, 2009; Benmamoun *et al.*, 2013; Debenport, 2011; Rothman & Treffers-Daller, 2014; Stern, 1983), ambiguity of the terms, which can impede advancement of rigorous theory and methodology (Cheng *et al.*, 2021), and misrepresentation of NS as good or ideal speakers, and NNS as bad or abnormal speakers (Vulchanova *et al.*, 2022).

Here, we demonstrate that the concept of “nativeness” is often tied to culture-specific perspectives surrounding language use. We contribute to this special issue by carrying out a preregistered study that highlights how increasing participant diversity challenges the idea of what it means to define someone as a NS or NNS. We show how those terms are often inherently determined by the social-cultural values of a community and therefore they simultaneously serve and limit the advancement of psycholinguistics research.

Linguistic stereotyping of NNS is common and easily manipulated

Although an accent is a common feature of NNS (Moyer, 2013), it can result in considerable stereotyping by the listener. These stereotypes can also have detrimental effects on speakers’ credibility (De Meo, 2012), which can influence immigration status, courtroom proceedings, and even job hiring practices (Smith, 2005), for example, NS are more qualified to teach English than NNS (Holliday, 2006). Even a medical doctor with an accent can be considered less competent than a doctor without an accent (Baquiran & Nicoladis, 2020). When compared with standard speakers, nonstandard speakers are dispreferred for high status employment, discrimination that increases with the strength of a speaker’s nonstandard accent (Carlson & McHenry, 2006). While such attitudes negatively affect speakers across low prestige varieties of a language, foreign accents tend to be downgraded the most consistently (Dragojevic *et al.*, 2021).

Dragojevic and Goatley-Soan (2022) found that variance in language attitudes can create a linguistic hierarchy depending on listener-perceived prestige of foreign accents. In their study, 245 US residents judged varieties of English including standard and nonstandard American English that they identified as belonging to various groups (e.g., Hispanic, French, German, Russian, Arabic, Farsi, Hindi, Mandarin, Vietnamese). Listeners were asked to rate linguistic forms in terms of *status* and *solidarity*. Although all nonstandard speech was rated lower than standard speech, some foreign accents (e.g., Arabic, Farsi, and Vietnamese) were more prone to prejudice than other accents (e.g., French and German).

Bias judgments may also occur as a result of listeners’ prior experience in regard to NNS (e.g., Lindemann, 2003; Hu & Lindemann, 2009; Kang & Rubin, 2009). Sheppard *et al.* (2017) asked instructors who teach content courses and instructors

who teach language skills to international students to rate L2 speech in terms of comprehensibility and intelligibility. The authors found no difference overall, but did find that instructors who teach content courses, that is, those with less experience with NNS, showed a correlation such that those with negative perceptions about the linguistic abilities of international students gave lower comprehensibility ratings than those with positive perceptions.

Given the well-established finding that listeners have stereotypes and biases toward NNS (e.g., Dragojevich & Goatley-Soan, 2022; Hu & Lindemann, 2009; Kang & Rubin, 2009; Ramjattan, 2019; Rubin, 1992; Sheppard et al., 2017), a growing body of research has asked whether these biases can be manipulated. Reid et al. (2019)—the study we conceptually replicated here—tested whether listeners can be positively or negatively biased toward NNS through a short interaction with an experimenter. In the positive condition, the experimenter tells a brief anecdote about her positive experience at a local cafe, where she was served by a French NS with “excellent” English skills. Conversely, in the negative condition, the experimenter criticizes a French NS’s English. This casual interaction between the participant and the experimenter took place prior to a speech rating task. In the task, listeners heard L2 English speech from Québécois French NS and rated the speech in terms of accentedness, comprehensibility, segmental errors, intonation, and flow on a scale of 0 to 1000. Listeners were biased positively and negatively compared to a control (no manipulation) group. Listener age also affected the results: younger learners manipulated with positive bias tended to be more lenient (for accentedness, comprehensibility, intonation, and flow) than control listeners. The same was true for older listeners but only when they measured comprehensibility and intonation. Interestingly, the effect of negative bias resulted in a different pattern. Specifically, younger listeners continued to be more lenient in all five measures even after hearing a negative statement, while the older listeners did not show similar favoritism, but rated speech lower than the control group in all five features.

In a follow-up study, Reid et al. (2020) examined whether teachers of German can be manipulated to change their bias toward non-native German speech. Specifically, teachers of German were asked to rate non-native speech for accentedness, comprehensibility, segmental errors, intonation, and flow. The authors also examined whether teachers reacted differently to the manipulation given their own NS/NNS status. That is, half of the teachers were regarded as native German teachers (born to German-speaking parents and learned the language before school) and half were non-native, that is, from non-German families and with no or late exposure to German. The researcher, while setting up the experiment, complained about German-majoring students’ inadequate grammar and accent. This negative comment was delivered to one-half of the teachers prior to their ratings of non-native German speech. Another half of the teachers did not hear any comments, that is, the control condition. Results revealed that in the control condition, NS/NNS teachers judged comprehension, accentedness, and segmental errors differently: native teachers demonstrated more leniency than non-native teachers. However, native teachers were more susceptible to the negative bias, rating more harshly the same features (i.e., comprehension, accentedness, and segmental errors) that diverged from non-native teachers’ ratings in no-manipulation condition. All teachers, regardless of NS/NNS status, upgraded flow and intonation ratings when exposed to negative bias.

In another follow-up, Reid et al. (2021) examined whether listeners’ social biases could be reduced through task practice. In this study, English–French bilinguals were

asked to measure non-native speech for comprehensibility and accentedness. Prior to hearing bias-stimulating (either negative or positive) anecdotes and rating non-native speech, half of the listeners were asked to complete a similar speaking task either in their dominant language (English) or less dominant language (French). The authors hypothesized that a shared experience might reduce the impact of social bias. Specifically, they predicted that task practice in English compared to practice in French would reduce naive listeners' tendency to overrate non-native speech. The authors found only negative (not positive) priming resulted in a statistically significant difference with a control group, and only in judging accentedness (not comprehension). That is, listeners, when exposed to either positive or negative social bias, generally showed solidarity by perceiving non-native speech to be more comprehensive and accent-free. Listeners could reduce (i.e., match with the control group) their social bias only when provided with a negative statement and asked to complete the task in English (their dominant language). This was true both for L2 comprehensibility and accentedness judgments.

Kutlu *et al.* (2020, 2022) demonstrated that biases toward NNS can be shaped by their geographical locations (e.g., Gainesville, Québec) and their social network (measured in terms of exposure to the same or other racial and ethnic groups). In Kutlu *et al.* (2020), the authors attempt to determine whether a listener's social network diversity and seeing a speaker's picture (i.e., Caucasian and South Asian faces) impacted their perceptions of American, British, and Indian English speakers. Within this study, there were 58 listeners across different races, all of whom were native speakers of American English. They were required to complete a language background questionnaire, English proficiency task, social network questionnaire, and rate speech for intelligibility and accentedness. The speakers in the study were six Indian English speakers, six female speakers of British English, and two female speakers of American English. For the intelligibility task, listeners viewed the speaker's face, listened to the speech, and typed sentences based on what they have heard. Similarly, the accentedness task required participants to listen to the sentences once more and use a 9-point Likert scale in determining if the speaker had an accent by selecting number buttons on the keyboard.

Kutlu *et al.* (2020) found an interaction between a speaker's face, speech varieties, and the listener's racial and social background. There was a significant difference in intelligibility and accentedness judgment given which face was displayed, with South Asian faces being rated as less intelligible and more accented. Additionally, Indian English when paired with a Caucasian face received higher intelligibility scores than when paired with a South Asian face. Similarly, a judgment of American English as the most accented was pronounced when heard with a South Asian face. Lastly, it was found that listeners with more diverse racial exposure showed less "bias" in judging accentedness, although no difference was detected in how they perceived intelligibility. These perceptions of accentedness and intelligibility are not merely a result of what listeners hear, but an indication of societal perceptions of language and race.

The current study: Preregistered conceptual replication and extension of Reid *et al.* (2019)

Reid *et al.* (2019) showed that "native" listeners could be easily biased toward "non-native" speech across five linguistic dimensions. The negative social bias

manipulation involved an English NS (the researcher) who felt they were not served “adequately in English by a native French-speaking [restaurant] employee” who had “an atrocious accent . . . poor grammar, and had not bothered to learn the other official language of Canada (pp. 426–7).” The concept of a “native” speaker in Montreal most likely carries very specific connotations. In Montreal, upwards of one-third of the population identifies as a visible minority, the majority of whom are Black (9.1% of the total population) and Arab (6.4%).¹ The content of the social bias manipulation probes the racialized discourse of who can be considered a legitimate speaker of English. In Reid et al.’s (2019) literature review, many studies are cited that support this race-based bias (image-based manipulation in Rubin, 1992; racial category of the speaker implied by the researcher in Hu and Lindemann, 2009). While the manipulation was not explicitly based on racial categories in Reid et al. (2019), the demographic context of Montreal and racialization of L2 English speakers is still influential in perceptions of their speech. Moreover, Reid et al. (2019) described “ethnic language” and ethno-national labels (“Anglophone Québécois” vs. “Québécois”) as background characteristics but did not extend the potential influence of the raters’ racialized identities any further.

We problematize a series of givens, as Pennycook has called for in our field (2001, p. 7), the first of which is that L2 speech ratings are not influenced by the race of the listener. There are (at least) two layers to this claim: the racialization of both the speaker and the listener. Available cross-linguistic and cross-cultural data indicate that speech ratings are not influenced by the phenotypic features or ethno-cultural origins of listeners. However, how individuals have been racialized in the societies they have existed in may influence how they perceive the speech of others. In other words, while the *race* of speech raters is not meaningful, *racialization* may be. When the target language of evaluation is a language that has been both globalized and localized at the same time (English), the potential for *membership* and *inclusion* can be fraught for both the L1 and L2 speaker. Therefore, asking someone to 1) self-identify as a *native* speaker and then 2) *As a NS*, rate other speakers, is a complex request in its racialized considerations. The need for such problematization of NS ideology exists in every language, yet most urgently for English.

Here, we examine “native” speech perception biases across a wide range of varied (self-claimed) “native” English speakers. Informed by Porte and McManus’ (2018) call for “modified replications” and Marsden et al.’s (2018) systematic review of replication studies, we frame our study as a conceptual replication and extension. We chose Reid et al. (2019) for the following four reasons: 1) thematic importance to the field (perceptions of L2 speech are not objective and standardized, nor insulated from bias), 2) recency and impact on the field (19 citations since 2019, count provided by Cambridge University Press), 3) replicability (the study’s authors made their materials available on IRIS, facilitating a replication), and 4) attempt to include participants from racially diverse communities in order to generate more reliable understanding of human behaviors. We strictly adhered to the following methodological decisions in the initial study: a) speech materials, b) speech sample rating categories and descriptions, and c) rating instructions and scale. The initial study’s authors made these materials available on IRIS and thus we were able to use their exact materials.

We chose to change a number of other method and procedure details as follows: (a) number of participants, (b) geographical background of participants, (c) online

format of experiment, (d) environment in which the experiment took place, (e) sample task speech samples, (f) background and social attitudes questionnaire, and (g) presentation of the bias. We increased (a) for greater statistical power and expanded online recruitment (b) to any English-speaking country in the world for greater generalizability. We achieved a geographical distribution of 12 different countries all of which consider English an “official language” and/or language of education (United Kingdom, US, Canada, Scotland, Ireland, Australia, South Africa, Nigeria, Zimbabwe, Singapore, Malaysia, and Hong Kong) and even representation among three racial categories. In response to the COVID-19 pandemic, we tested participants online, which resulted in changes (c) and (d) and (g). Because the initial study’s practice speech samples (prior to the main task) were not available, we generated these samples (e). Whereas we did not alter the questionnaires (f) to include items about social group indicators, the wording in and the omission of some items occurred only when the initial study’s local context was not applicable (items specific to French, Québec, Montreal, and Canada).

In addition to the intentional, motivated changes detailed above, there was one unintentional change to the initial study design which was not specifically motivated: In our study, we presented the background questionnaire and social attitudes questionnaire before the listening task, whereas Reid *et al.* presented them at the end. We acknowledge that the change in order may have influenced ratings. Although it is an empirical question whether social attitude reflection caused listeners to change their ratings, we ran multiple exploratory tests using answers from the social attitude questionnaire to predict behavior and found no significant predictors (see online R code).

To summarize, our conceptual replication of Reid *et al.* (2019) investigates whether social biases, manipulated by exposing native English speakers to negative or positive comments about NNS language abilities, impact their ratings of non-native speech across five linguistic dimensions. The study’s two research questions are:

1. Does a social bias orientation (positive or negative) influence “native” listeners’ ratings of “non-native” English speech when testing a wide range of diverse “native” listeners via the internet?
2. Does a social bias orientation (positive or negative) influence to the same degree self-identifying Asian, Black, and Caucasian “native” listeners?

Expanding the population from native listeners from one community (Montreal) to native listeners from diverse geographical backgrounds through the English-speaking world may change participants’ sensitivity to social bias in evaluating non-native speech given that the “native” and “non-native” labels are social-cultural creations. As a social bias finds a foothold consistent with contexts of similar social assumptions and constructs (e.g., anglophone speakers in a majority-French context of Montreal in Reid *et al.*), delocalizing the social experience of listeners should weaken any effect of bias. Moreover, as social constructs of race have determined unequal claim to “nativeness,” we invite our participants to self-identify as NS and expect that a social bias orientation may vary considerably as a function of listeners’ race, geographic location, and age (Kutlu *et al.*, 2020; Reid *et al.*, 2019, 2020). For these reasons, an internet-sampled population containing a more balanced distribution of participants across races, locations, and ages may not yield an effect of social bias.

Method

Our study was preregistered on the Open Science Framework. All methods follow our registered report except where noted. All research materials, R analysis code, and data are available on the [Open Science Framework](#). The study was approved by the authors' Institutional Review Board. All participants were paid for their time.

Positionality statement

We are a team with different ethno-racialized, linguistic, and research backgrounds. Our research team includes members who identify as an Asian-American NS of English, two Asian NNS of English, an Afro-Caribbean speaker of English, and two Caucasian American NS of English. Not only do we speak different first languages (e.g., Taiwan Mandarin Chinese, American English, Bahamian Creole English, Kazakh, and Russian) but we also speak multiple additional languages (e.g., Taiwanese, English, French, Italian, Korean, Latin, Spanish, Turkish, Mandarin) with varying degrees of proficiency. As noted in Cheng et al. (2021), because the word *native* in Russian is linked to “national identity from its association with states” (p. 10), speakers of Russian from post-Soviet areas may hesitate to regard themselves as NS. This was a familiar case for one of the authors of this study representing the “non-Russian” Russian-speaking community. Our sensitivity to such nuances impacted our methodological decisions specifically in how we set the Prolific filtering by using the following specifications: “first language,” “primary language,” “fluent language,” and how we geographically expanded a pool of the target population. In addition to our research practice, most of us have taught languages we acquired as adults, identifying as NNS teachers of these languages. Informed by such experiences, we recognize the problems with a native/non-native binary categorization. Furthermore, we believe that these challenges and insecurities we had as “non-native” teachers and multilingual speakers have been heavily influenced by the literature and research we were exposed to while growing as researchers in linguistics and psycholinguistics. These factors affect our work in many critical ways, reflected in the choice of study we replicated, in the decision to add racial listener categories as a novel contribution to the repetition, and in how we narrate and interpret our study results.

Moreover, we have diverse research experiences and interests which we believe complement our strengths and help to discover areas for improvement. For example, half of the contributors were trained as qualitative researchers with backgrounds in linguistics, investment, and identity, while others practiced predominantly quantitative research methodology with a focus in applied linguistics and psycholinguistics. This combination informed how we viewed the results (e.g., attitude toward statistically non-significant output) and interpreted interesting findings (e.g., pride in ethnicity by Black group). Overall, the shift toward challenging the notion of “nativeness” is timely as we seek to explore if a listener's social bias orientation influences individuals of various racialized identities in the same manner.

Participants

In our registered report, we planned to recruit 288 participants self-identifying from four racial groups: Asian, Black, Caucasian, and mixed-race. However, due to

increased payments to account for the longer than expected time on task, we went over our budget and were only able to collect data from 216 participants across three groups. Out of the four categories, we decided to exclude the mixed-race group since it could represent any combination of the target groups. Given more resources, we would like to include a mixed-race group in future studies. Racial categories are imprecise and socially constructed—this is partly why we included this variable in the study—and are thus understood differently around the world. All three racial categories were based upon self-identification, without any qualifiers. As our participant pool spanned the entire world—anywhere there was a stable internet connection and access to our recruitment site and experiment platform—denominating sub-categories to qualify each macro-category could potentially introduce further confusion.

The 216 raters were recruited via the online recruitment platform, Prolific. All participants we report on had completed at least 60 studies on Prolific (prior to our study) and had a Prolific approval rate of 98% or higher. Participants needed to satisfy three criteria to be classified as a “native” English speaker for our study: 1) their first language acquired was English; 2) their primary language they used was English; and 3) they self-identified as a fluent English speaker. Responses to these questions were set when the user created their Prolific account (and cannot be changed once an account is created) and are therefore all self-identified. In other words, we did not rely on assessments of proficiency to select participants. In an effort not to restrict the sample to a prescriptivist, top-down approach to non-expert NS evaluation of L2 speech, our sample had the possibility for greater diversity. For example, traditional interpretations of a NS may assume a speaker is from what is referred to as an “inner circle” country such as the US or the UK. “Outer” or “expanding circle” countries where English is widely spoken or is an official language such as India and Nigeria are frequently not included. Our criteria do not make any assumptions.

All participants were 18 years old or older, had no self-reported history of hearing impairments, and self-identified as belonging to one of the following racial groups: Asian ($n = 72$), Black ($n = 72$), and Caucasian ($n = 72$). Of the 216 participants, 27 were removed for the following issues: failing the bot check ($n = 3$), taking over 90 min to complete the listening task ($n = 20$), or finishing the listening task in under 10 min ($n = 4$). This left a total of 189 listeners, including 63 participants of three different ethnicities, respectively.

All participants completed a background and social attitudes questionnaire following Reid *et al.*'s (2019) format. In Reid *et al.*, all 60 listeners self-identified as coming from monolingual households with English ($n = 51$) reporting to be their primary language. However, in our sample one-third of participants ($n = 67$) indicated to be proficient in languages other than English, although their daily use and communication with non-native English speakers were low 8% and 15%, respectively. Although self-reporting language proficiency introduces considerable variation, in order to closely follow Reid *et al.* (2019), we modeled our background questionnaire after theirs. Reid *et al.* did not use any objective measures other than for the language of schooling, and if the respondent had taken any linguistics classes. All other measures are percentages of time using English, using other languages, and interacting with NS of both English and languages other than English. Table 1 reports a summary of the participants' language background and social attitude questionnaires.

Our three groups differed in four measurements: *daily use of English*, [$F(2,180) = 3.62, p = .03, \eta_p^2 = .04$]; *daily use of English with native speakers*, [$F(2,180) = 8.37, p < .001, \eta_p^2 = .08$]; *listening to English media*, [$F(2,180) = 6.95, p = .001, \eta_p^2 = .07$]; and *pride in ethnic group*, [$F(2,180) = 17.47, p < .001, \eta_p^2 = .16$]. All other comparisons of measurements were null ($ps > .05$). Figure 1 shows individual participant responses, box plots, and density plots for the four measurements in which group differences were found. In Table 1, the rating for daily speaking of English, daily use of English with NS, daily listening to English media, daily use of other language, and daily use of other language with NS are based on a 0–100% scale. The last three questions are based on the Social Attitudes Questionnaire (see OSF materials). The maximum total points for the question “pride in my ethnic group” and “feeling toward other ethnic groups” is 45, and for the question “attitudes toward immigrants” is 36. The differences observed will be explored further in the discussion section.

Speech materials

We used Isaacs and Trofimovich’s (2012) recordings of 40 Québécois L1 French speakers of English as an L2 (the same as used by Reid et al.). These recordings included 27 women and 13 men, all French NS who were born and grew up in French-speaking households in Québec. They were educated entirely in French, and their ages ranged from 18 to 61 ($M = 35.6$). This audio contained the first 30 s of the “Suitcase Story,” a picture prompt that each speaker narrated freely (see Derwing et al., 2004). Before beginning the main rating task, listeners practiced rating with three original audio samples recorded by three L2 English speakers of different L1s (Mandarin Chinese, Kazakh, and Italian, respectively). These practice speech samples included recordings by two contributing authors. As in the initial study, these speakers summarized the “Suitcase Story” in a free-flowing narrative while viewing the same series of pictures. The speakers recorded the audio themselves on their own computers.

Rating procedure

Participants logged onto the experiment hosted on the Gorilla platform (Anwyl-Irvine et al., 2020). After consenting to participate in the study, participants were asked to wear headphones and confirm that they would do so for the remainder of the study. Next, participants completed the background questionnaire and social attitudes questionnaire.

Before beginning the main listening task, a bot check confirmed that the participant was paying attention. Participants were then given instructions about the rating task following Reid et al.’s design. Participants were shown the “Suitcase Story” picture sequence and then given instructions for each rating category that included definitions and examples. They then completed three practice ratings with unique audio samples. In each trial, listeners clicked on an audio button to initiate the 30-s sample and moved a sliding scale from 0 to 1,000 for the first two variables: 1) accentedness and 2) comprehensibility. The scale endpoints were indicated with qualitative descriptions (e.g., “heavily accented” near 0, and “no accent at all” near 1,000). This scale was modeled after that used in Reid et al. (2019, p. 426), and the

Table 1. Listener’s Language Background and Social Attitude Characteristics

Background variable	Bias	Asian		Black		White	
		<i>M (SD)</i>	Range	<i>M (SD)</i>	Range	<i>M (SD)</i>	Range
Age	POS	32.75 (11.25)	18–57	33.30 (12.18)	19–64	37.29 (13.38)	21–66
	NEG	32.09 (10.39)	19–53	33.71 (13.10)	20–68	35.76 (13.65)	19–61
	CTRL	32.09 (10.39)	19–62	35.09 (14.21)	19–77	35.76 (14.37)	21–68
Daily speaking of English	POS	89.50 (14.32)	60–100	97.50 (5.50)	80–100	96.19 (8.26)	60–100
	NEG	91.82 (12.20)	50–100	92.38 (11.79)	60–100	97.14 (7.17)	80–100
	CTRL	94.29 (11.21)	50–100	95.91 (6.66)	80–100	96.19 (9.74)	80–100
Daily use of English with native speakers	POS	77.00 (25.36)	10–100	96.00 (9.40)	60–100	87.62 (11.36)	70–100
	NEG	79.09 (20.91)	20–100	84.76 (24.62)	0–100	94.29 (11.65)	50–100
	CTRL	84.29 (22.93)	10–100	94.55 (5.96)	80–100	90.48 (15.32)	40–100
Daily listening to English media	POS	89.50 (10.99)	60–100	97.50 (5.50)	80–100	96.67 (9.13)	60–100
	NEG	91.36 (14.57)	50–100	94.76 (6.80)	80–100	96.67 (5.77)	80–100
	CTRL	90.48 (19.36)	10–100	98.64 (4.68)	80–100	94.76 (9.28)	70–100
Daily use of other language	POS	14.00 (17.89)	0–60	4.50 (10.99)	0–40	2.86 (7.17)	0–30
	NEG	6.82 (9.45)	0–30	11.90 (21.59)	0–90	10.95 (25.08)	0–100
	CTRL	5.71 (11.21)	0–50	9.09 (16.01)	0–70	4.76 (8.73)	0–30
Daily use of other language with native speakers	POS	33.00 (40.92)	0–100	12.00 (26.68)	0–90	4.76 (10.78)	0–40
	NEG	19.09 (31.76)	0–100	19.05 (33.30)	0–90	11.90 (27.13)	0–100
	CTRL	10.00 (22.58)	0–100	10.00 (21.38)	0–100	14.29 (32.49)	0–100
Pride in my ethnic group	POS	32.25 (10.25)	05–45	39.80 (8.59)	16–45	31.52 (8.59)	19–45
	NEG	32.59 (8.73)	15–45	40.14 (7.72)	18–45	33.71 (8.38)	21–45
	CTRL	34.43 (7.54)	21–45	40.32 (7.86)	16–45	28.19 (11.92)	08–45
Attitudes toward immigrants	POS	14.70 (6.38)	04–26	14.15 (8.22)	05–36	14.00 (7.20)	04–27
	NEG	15.09 (6.25)	04–27	15.76 (7.50)	05–34	16.00 (7.46)	07–29
	CTRL	17.81 (8.66)	04–36	12.32 (5.49)	04–21	15.29 (10.52)	04–35
Feeling toward other ethnic groups	POS	34.80 (6.95)	14–45	31.70 (6.23)	17–45	35.57 (5.10)	26–45
	NEG	37.50 (6.27)	25–45	34.48 (6.85)	19–45	34.71 (6.02)	21–44
	CTRL	33.90 (6.83)	17–45	33.23 (5.74)	21–44	32.76 (6.33)	20–43

Note. POS refers to positive bias, NEG refers to negative bias, and CTRL refers to baseline control without any bias manipulation.

following features were the same: 1–1,000 scale; no numeric endpoints; and no interval markings. In the absence of marked numeric intervals, listeners could see the exact number rating as they moved their marker along the sliding scale. Listeners had to click on the audio button and respond to both ratings before being

able to advance to the next screen and were only permitted to listen to the audio one time. Participants listened to the recordings according to the procedure used in the initial study. The explanation Reid et al. provided for limiting the recording to be played once when evaluating the speaker's comprehensibility and accentedness was "the assumption that accent and comprehensibility reflect initial, intuitive perceptual judgments" (2019, p. 426). On the next screen, listeners were able to replay the audio as needed and rated the remaining variables of 3) variable and consonant errors, 4) intonation, and 5) flow. The rating categories and instructions are taken from the initial study and center the listener's perspective in generating ratings. For example, for flow: "Speakers can speak at a natural rate and can be comfortable to listen to"; for intonation: "Intonation should come across as natural and unforced"; and for comprehensibility: "If you can understand with ease, then a speaker is highly comprehensible." The sliding scale and requirement to play the audio and complete the ratings before advancing to the next audio sample were the same on this screen as on the previous screen. A progress bar labeled "rating task percentage completed" was updated as the listeners completed each of the 40 audio samples, and at any time the listeners could navigate back to the detailed instructions presented before the practice tasks to review the rating category definitions.

After completing the practice sessions but before starting the main 40 trials, we presented our bias. Whereas, in Reid et al., the lab setting allowed for incorporation of manipulated bias in the form of an anecdote casually shared by the experimenter to the participant, we presented our manipulation via audio recording before the practice questions and directly addressed the participants with statements about non-native English speakers instead of the original Canadian-specific situation. The original social bias stimuli (and social attitude questionnaire) focused heavily on Canadian sociopolitical contexts (i.e., social status of English in Québec). Therefore, the presentation of our social bias stimuli was void of nation-specific references. The positive and negative stimuli were about 40 s in length, recorded by a Caucasian American male English NS, and recounted the need to improve grammar and accent to be more "native-like" (negative bias):

Since you are a native English speaker, you can perceive differences in how well non-native speakers speak English. For example, when you go to your local grocery store, you can clearly tell if the cashier doesn't speak English as their first language. Even then, you can tell when they don't put very much effort into trying to sound like a proper native speaker of English. You can hear a very distinct accent influenced by their first language, and sometimes their grammar doesn't even make any sense. Anyone who moves here should be able to speak it fluently, especially if they plan on getting a job where they have to interact with people in English!

Or praising the multilingual skills of NNS (positive bias):

Since you are a native English speaker, you can perceive differences in how well non-native speakers speak English. For example, when you go to your local grocery store, you can clearly tell if the cashier doesn't speak English as their first language. However, you also know that they put in a lot of effort trying to

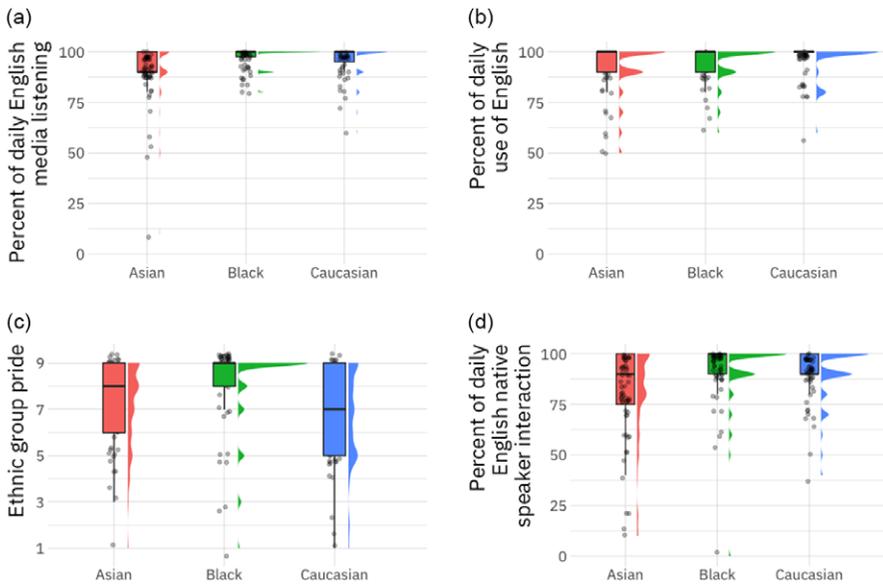


Figure 1. Self-reported background questionnaire measures in which group differences were found: percent of daily English media listening (a), percent of daily English use (b), ethnic group pride (c), and percent of daily English native speaker interaction (d). Note that a, b, c, and d all have different y-axis scales.

become a fluent English speaker, since English has so many rules and exceptions. Sometimes you can hear a slight accent from their first language, but usually their grammar is spot-on, even if they do make a mistake or two. It's really impressive that they're fluent enough in English to use it for work purposes when they probably grew up speaking a totally different language!

The baseline condition presented only the brief audio clip presented to all participants thanking listeners for participating in the study.

Thank you for participating in this study. Your task will involve rating a series of audio samples for English fluency.

After the bias was played, the experiment automatically advanced to the instructions. The rating task had a time limit of 90 min. Gender did not enter into the randomization of the presentation of the speech files. Following the rating task, participants answered a debrief questionnaire containing four sliding-scale questions. In total, the experiment lasted approximately 40 min. Table 2 reports the number of participants in each condition.

Data analysis

Cronbach's alpha was calculated to check for internal consistency across listener groups and bias conditions and showed high reliability ranging from .86 to .97 (see R code). Next, to ensure that the participants were not aware of the manipulation, we examined

debrief questionnaire responses on whether the experience was pleasant, rating was difficult, and how confident they were in their ratings. We followed the same measurement scale as in Reid et al. (2019). Results revealed that there was no significant difference in how participants rated their pleasantness across the session, the helpfulness of the instructions, and difficulty of the rating ($ps > .05$). However, the main effect of race was significant for the question on rating confidence, [$F(2,174) = 3.13, p = .04, \eta_p^2 = .04$]. The Asian group ($M = 67.21, SE = 2.76$) reported feeling less confident in their ratings compared to the Caucasian ($M = 76.60, SE = 2.76$) and the Black ($M = 74.92, SE = 2.90$) groups; further post-hoc analysis revealed that the difference was statistically significant only between the Asian and the Caucasian groups.

To examine the effect of social bias on accentedness, comprehensibility, segmental errors, intonation, and flow of non-native speech, five separate multilevel regression models were run using the lme4 package (Bates et al., 2015) and the lmerTest package (Kuznetsova et al., 2017) in R version 4.1.0. Bias was treated as a 3-level categorical variable with the “no bias” condition serving as the reference level. This allowed for comparisons of no bias-positive and no bias-negative. Bias was also included as a random item slope. Race was treated as a 3-level categorical variable with Caucasian as the reference level. This allowed for two comparisons: Caucasian-Black and Caucasian-Asian. (The Black-Asian comparison was obtained using estimated marginal means from the emmeans package; Lenth et al., 2019). Age was included as a continuous factor and random participant slope. For each model, all two-way and three-way interactions were tested. In all five models, all two-way interactions and the three-way interaction resulted in singular models with variance inflation factors greater than 10 and were therefore removed from the model. The final model tested contained no interactions:

dependent variable \sim bias + race + age + (bias | item) + (age | participant).

Results

The results from multilevel modeling investigating the effect of social bias manipulation on five dimensions of L2 speech revealed no significant effects in any of the five models. All interactions were null at an alpha level of .05. Figure 2 plots the five models' estimates with the coefficient on the x-axis and the model term on the y-axis. For accent, there was a null effect of bias manipulation (positive: $\beta = -21.38, SE = 19.13, p = .27$; negative: $\beta = -9.07, SE = 19.00, p = .63$), race (Black: $\beta = 15.61, SE = 19.29, p = .42$; Asian: $\beta = 30.53, SE = 19.23, p = .11$), and age ($\beta = -0.63, SE = 73.98, p = .39$).

For comprehensibility, there was a null effect of bias manipulation (positive: $\beta = -2.32, SE = 22.31, p = .92$; negative: $\beta = -26.54, SE = 22.17, p = .23$). Furthermore, there was no effect of race (Black: $\beta = 9.98, SE = 22.27, p = .66$; Asian: $\beta = 4.69, SE = 22.51, p = .84$) or age ($\beta = -0.72, SE = 0.75, p = .34$).

For segments, there was a null effect of bias manipulation (positive: $\beta = -3.12, SE = 18.88, p = .87$; negative: $\beta = 17.25, SE = 19.16, p = .37$). Similarly, there was a null effect of race (Black: $\beta = 1.10, SE = 19.09, p = .95$; Asian: $\beta = 5.76, SE = 19.04, p = .76$) and age ($\beta = -0.97, SE = 19.09, p = .18$).

Table 2. Number of Participants in Each Condition

Race	Social bias			Total
	Negative	Positive	Control	
Asian	22 (9)	20 (9)	21 (11)	63 (29)
Black	21 (8)	20 (7)	22 (6)	63 (21)
White	21 (7)	21 (7)	21 (9)	63 (23)
Total	64 (24)	61 (23)	64 (26)	189

Note. The number in parenthesis is the number of male participants per condition.

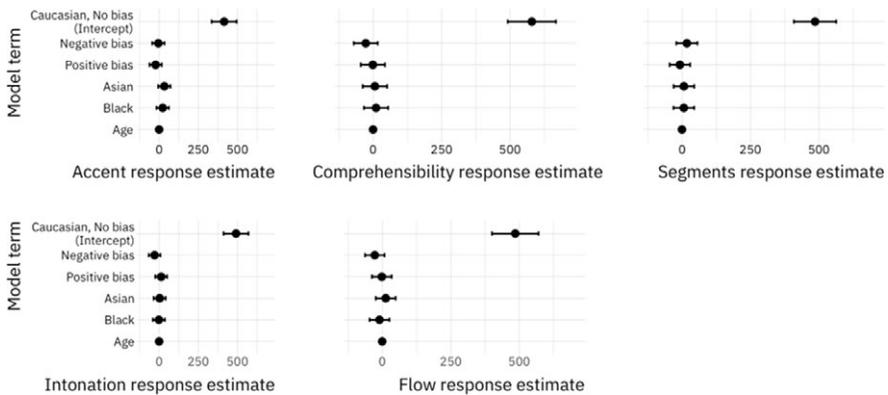


Figure 2. Regression model coefficients for the five responses to speech.

For intonation, there was a null effect of manipulation (positive: $\beta = 14.07$, $SE = 19.21$, $p = .47$; negative: $\beta = -29.44$, $SE = 18.96$, $p = .12$). There was no significant effect of race (Black: $\beta = -4.25$, $SE = 19.37$, $p = .83$; Asian: $\beta = 1.82$, $SE = 19.38$, $p = .93$) or age ($\beta = -0.51$, $SE = 0.70$, $p = .47$).

Finally, for flow, there was no significant effect of bias manipulation (positive: $\beta = -0.11$, $SE = 18.02$, $p = .99$; negative: $\beta = -28.55$, $SE = 17.82$, $p = .11$). There was no significant effect of race (Black: $\beta = -15.95$, $SE = 18.08$, $p = .38$; Asian: $\beta = 10.20$, $SE = 18.09$, $p = .57$) or age ($\beta = -1.01$, $SE = 0.66$, $p = .13$).

Discussion

With respect to our first research question, we failed to replicate Reid *et al.* (2019). We found no effect of social bias in our study and no effect of age. Regarding our second research question, we found that listeners who self-identify as Asian, Black, and Caucasian showed no difference in their behavior. Moreover, there was no effect of age and no interaction among the three predictors. Given previous research that indicated the listener’s background can affect speaker judgments (e.g., Kutlu *et al.*, 2020, 2022; Kang & Yaw, 2021), we suggest at least three accounts for our null results.

First, by increasing participant diversity, we demonstrated that the NS/NNS terms became more variable presumably because we tapped into different cultural perspectives surrounding language use. Our sample of native listeners was not as “native” as is commonly defined in the literature (see Cheng et al., 2021) and did not resemble monolingual-like listeners in Reid et al. (2019) within a specific environment. We recruited listeners on Prolific using common, self-identified filtering requirements in psycholinguistics studies. Native listeners in our study were more heterogeneous in their linguistic backgrounds than in Reid et al.; one-third of participants ($n = 67$) reported to be proficient in other languages in addition to English, of which 22 also filled out multiple languages as their “native” language. Additionally, in each racial group, at least one participant indicated a language other than English as their sole native language in our survey (we note this suggests participants may have lied on the Prolific account in order to be eligible for more studies). This includes Asian (Cantonese [$n = 2$], Tagalog, Malay), Black (Igbo, Shona, Zulu, Xitsonga, Setswana), and Caucasian (Croatian). This implies that NS who consider themselves to be fluent speakers of English whose first language learned was English do not always regard English as their “native” language, thus having a clearly different concept of nativeness than what has been commonly practiced in the field. This observation substantiates the argument that the notion of “nativeness” should not be attributed merely to monolingualism, or to English speakers in Anglophone countries, but requires reconceptualization depending on “which aspect of language experience [researchers] are investigating” (Cheng et al., 2021, p. 18).

This idea of “nativeness” was challenged by the geographic diversity of our participants with respect to Reid et al. (2019). Instead of residents of Montreal, born and raised in Québec in monolingual English-speaking households, our listeners reported residence in 12 different countries. If a bias is ascribed to socially fueled stereotypes such as listener’s background and subsequent expectations (see Kang & Rubin, 2009 on reverse linguistic stereotyping), the degree to which such a bias is projected can vary based on language ideology/policy of a specific geographical context. Reid et al.’s effect could be closely tied to the Montreal, Canada setting, which is a unique environment for research on bilingualism and language contact (e.g., Fowler et al., 2008; Tiv et al., 2020). Related to this, Kutlu et al. (2022) highlighted how English–French bilinguals compared to English–Spanish bilinguals judged a Caucasian face with British English to be more accented while the Gainesville listeners judged Indian English with South Asian faces to be less intelligible. Clearly, the speaker/listener’s setting matters when discussing native/non-natives. Reid et al. (2019) added participants’ age as a construct because of a historical event (i.e., enactment of a French language policy in Québec). In this regard, our more global participant pool lacks a unified language-related event to have stoked a bias against (or for) a certain set of othered English NNS. As a result, participants who might have been able to identify the recordings as francophone L2 English speakers may have responded differently depending on what constitutes an “undesirable” non-native accent in their local contexts. Considering the political tension related to English and French language use in Montreal, we believe the effect Reid et al. found had less to do with “non-native” French-accented English or properties in the speech signal and more to do with a national-identity-related bias effect. We also noted that

in the initial study, nationality rather than race was used and listeners overwhelmingly identified with the label “Canadian” ($M = 8.3$, $range = 1-9$) over other labels such as French Canadian, Québécois, and “Other” where they had an open-response text blank. Where national discourses are concerned, NS ideology regarding inner and outer/expanding circle countries is a logical next step for exploration.

Second, the presentation format and methods of the two studies differed. There are many advantages associated with an online study design, most notably the diversity and inclusivity of the sample pool, which was a foundational motivation in our extension of Reid *et al.* Furthermore, asynchronous data collection expands accessibility to participation. We acknowledge that in-person testing has the advantage of keeping the participant on task and monitoring their behavior. It is possible that some of our participants became distracted during the task.

Third, the bias script and presentation differed between the original study and our study. While the mode of speaking in our recording was natural and conversational, it lacked the in-person contextualization of the initial study, delivered as a multi-tasking aside as the researcher set up for the study. In our study, the participants could not see the researcher and thus did not have the opportunity to develop any type of human connection, potentially limiting the authenticity of the bias. Instead, we contextualized the stimulus within the computerized context of the study, drawing a connection between the participant’s recruitment and their status of NS. In the debrief questionnaire, two questions specifically targeted the potential influence of the stimulus: “How helpful were the instructions during the session?” and the open-ended question, “Did any part of the study design affect your ratings?” Although participants did not specifically report on the bias stimuli, it is possible that some thought it artificial and/or were not paying attention as the recording played.

When questions are as complex as NS ideologies, there are certainly limitations of a survey instrument in capturing diverse perspectives. Qualitative methodologies could further probe these complexities, in particular regarding two interesting patterns that emerged in the data exploring ethnic pride and confidence ratings. The first pattern was observed in the social attitudes questionnaire, where Black listeners responded with a significantly higher rating to the statement, “I am proud to be a member of my ethnic group” than the Caucasian or Asian group did. (The ratings for the Caucasian and Asian groups were roughly equal [Caucasian $M = 31.14$, $SE = 1.13$; Asian $M = 33.09$, $SE = 1.13$]). While we offered an opportunity for participants to first self-identify in racial/ethnic membership categories that they themselves articulated (in an open-response text field), providing as examples hyphenated categories (e.g., “Chinese-American”) that moved beyond monolithic labels, we realize that pride in ethnicity is a fraught concept, particularly in 2022. The ethno-racial discourses in the US have become even more complex in the wake of the Black Lives Matter (BLM) and Stop AAPI Hate (Asian American Pacific Islander; see Liu, 2018). The geographic distribution of our participants creates an uneven mosaic of narratives regarding the acceptability of ethnic pride. Apartheid has framed ethnic identity in South Africa in a way that is not present in Malaysia, nor Nigeria. Furthermore, terminology in different countries (and within) regarding ethno-racial categories is not consistently applied. Morning (2015) specifies, “What is called ‘race’ in one country might be labeled ‘ethnicity’ in another, while ‘nationality’ means ancestry in some contexts and citizenship in others.” Importantly, the

intersection of nationalized conceptions of ethnic identity and linguistic identity has implications for NS ideologies.

The second pattern was seen in the debriefing questionnaire, where Asian participants rated themselves as significantly less confident in their ratings when compared with confidence ratings reported by Caucasian listeners. In a cross-cultural comparison study (e.g., Chen et al., 1995; Lee et al., 2002), Asian participants were found to be less inclined to choose extreme values in evaluation than their North American counterparts. For example, Lee et al. (2002) examined the cultural differences within the responses of a 13-item Likert scale questionnaire with Japanese, Chinese, and American participants, respectively. Both Chinese and Japanese participants (but not American participants) were more likely to select the midpoint of the Likert scale, rather than the endpoint for items pertinent to positive perceptions. This inclination could potentially transfer over into lower confidence in speech sample ratings, although this is also speculation.

Multilingualism may also contribute to confidence in ratings. In our post-hoc analysis, Asian listeners reported significantly lower daily use of English with NS and daily listening to English-language media when compared with both Caucasian and Black listeners. As these questions were asking about percentages compared with other languages used, it may have been that the Asian listeners were more actively utilizing multilingual practices on a daily basis than the other two ethnic groups.

Future research is needed in understanding how participants from different backgrounds evaluate one variety of language that they all self-identify to be NS of, and furthermore if there could be an effect of social bias in this situation. Since our conceptual replication reused the initial study's NNS recordings which all belong to the same ethno-linguistic group, the audio samples could be diversified to include L2 English speakers with different primary languages (as in Dragojevic & Goatley-Soan, 2022), and separate participants based on geographical region. This could uncover any variation in NS judgments of NNS from different linguistic backgrounds, as well as whether these biases are generalizable across the Anglophone world.

Considered together, our results demonstrate the fragility of a concept such as the "NS," as delocalizing a listener may deconstruct both the racial assumptions of nativeness and the social biases along NS/NNS fault lines. We conclude by joining Cheng et al.'s (2021) call for researchers and educators to avoid treating nativeness as a strict binary concept but rather to use specifications based on what aspects of language are of interest and relevant in their study, for example, exclusion/inclusion criteria, thus allowing for a more refined and accurate measure.

Replication Package. Replication data and materials for this article can be found at <https://osf.io/4wv9h/>.

Acknowledgements. We are very grateful for the feedback on earlier drafts of this manuscript. The Associate Editor, Kevin McManus, and the anonymous reviewers helped transform this paper for the better.

Conflict of Interest. Author Seth Wiener currently serves as an Associate Editor for *Applied Psycholinguistics* and played no role in the editorial process for this manuscript. The Editor-in-Chief invited an external and independent editor to handle the peer review process and make all editorial decisions.

Note

1 Montréal Population 2022. <https://worldpopulationreview.com/canadian-cities/montreal-population> accessed September 27, 2022.

References

- Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, *59*(2), 249–306. <https://doi.org/10.1111/j.1467-9922.2009.00507.x>
- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. <https://doi.org/10.3758/s13428-019-01237-x>
- Baese-Berk, M. M., McLaughlin, D. J., & McGowan, K. B. (2020). Perception of non-native speech. *Language and Linguistics Compass*, *14*(7), e12375. <https://doi.org/10.1111/lnc3.12375>
- Baese-Berk, M. M., & Morrill, T. H. (2019). Perceptual consequences of variability in native and non-native speech. *Phonetica*, *76*(2–3), 126–141. <https://doi.org/10.1159/000493981>
- Baquiran, C. L. C., & Nicoladis, E. (2020). A doctor's foreign accent affects perceptions of competence. *Health Communication*, *35*(6), 726–730. <https://doi.org/10.1080/10410236.2019.1584779>
- Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Benmamoun, E., Montrul, S., & Polinsky, M. (2013). Heritage languages and their speakers: Opportunities and challenges for linguistics. *Theoretical Linguistics*, *39*(3–4), 129–181. <https://doi.org/10.1515/tl-2013-0009>
- Carlson, H. K., & McHenry, M. A. (2006). Effect of accent and dialect on employability. *Journal of Employment Counseling*, *43*(2), 70–83. <https://doi.org/10.1002/j.2161-1920.2006.tb00008.x>
- Chen, C., Lee, S. Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, *6*(3), 170–175. <https://www.jstor.org/stable/40063010>
- Cheng, L. S., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., & Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology*, *12*, 1–22. <https://doi.org/10.3389/fpsyg.2021.715843>
- Cristia, A., Seidl, A., Vaughn, C., Schmale, R., Bradlow, A., & Floccia, C. (2012). Linguistic processing of accented speech across the lifespan. *Frontiers in Psychology*, *3*, 479. <https://doi.org/10.3389/fpsyg.2012.00479>
- De Meo, A. (2012). How credible is a non-native speaker? Prosody and surroundings. Methodological Perspectives on Second Language Prosody Papers from ML2P, 2012, 3-9. Retrieved from https://orientale.ricerca.unior.it/retrieve/handle/11574/40089/2649/De%20Meo_2012_how%20credible.pdf
- Debenport, E. (2011). As the Rez turns: anomalies within and beyond the boundaries of a Pueblo community. *American Indian Culture and Research Journal*, *35*, 87–110.
- Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*(4), 655–679. <https://doi.org/10.1111/j.1467-9922.2004.00282.x>
- Dragojevic, M., Fasoli, F., Cramer, J., & Rakić, T. (2021). Toward a century of language attitudes research: Looking back and moving forward. *Journal of Language and Social Psychology*, *40*(1), 60–79. <https://doi.org/10.1177/0261927X20966714>
- Dragojevic, M., & Goatley-Soan, S. (2022). Americans' attitudes toward foreign accents: evaluative hierarchies and underlying processes. *Journal of Multilingual and Multicultural Development*, *43*(2), 167–181.
- Fowler, C. A., Sramko, V., Ostry, D. J., Rowland, S. A., & Hallé, P. (2008). Cross language phonetic influences on the speech of French–English bilinguals. *Journal of Phonetics*, *36*(4), 649–663. <https://doi.org/10.1016/j.wocn.2008.04.001>
- Holliday, A. (2006). Native-speakerism. *ELT Journal*, *60*(4), 385–387. <https://doi.org/10.1093/elt/ccl030>
- Hu, G., & Lindemann, S. (2009). Stereotypes of Cantonese English, apparent native/non-native status, and their effect on non-native English speakers' perception. *Journal of Multilingual and Multicultural Development*, *30*, 253–269. <https://doi.org/10.1080/01434632.2020.1735402>
- Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*, 475–505. <https://doi.org/10.1017/S0272263112000150>

- Kang, O., & Rubin, D. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, *28*, 441–456.
- Kang, O., & Yaw, K. (2021). Social judgment of L2 accented speech stereotyping and its influential factors. *Journal of Multilingual and Multicultural Development*, Advance online publication. <https://doi.org/10.1080/01434632.2021.1931247>
- Kutlu, E., Tiv, M., Wulff, S., & Titone, D. (2020). The impact of race on speech perception and accentedness judgements in racially diverse and non-diverse groups. *Applied Linguistics*. <https://doi.org/10.1093/applin/amab072>
- Kutlu, E., Tiv, M., Wulff, S., & Titone, D. (2022). Does race impact speech perception? An account of accented speech in two different multilingual locales. *Cognitive Research: Principles and Implications*, *7*(1), 1–16. <https://doi.org/10.1186/s41235-022-00354-0>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in Nursing & Health*, *25*(4), 295–306.
- Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). Package ‘emmeans’.
- Lindemann, S. (2003). Koreans, Chinese or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics*, *7*(3), 348–364.
- Liu, W. (2018). Complicity and resistance: Asian American body politics in black lives matter. *Journal of Asian American Studies*, *21*(3), 421–451. <https://doi.org/10.1002/nur.10041>
- Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*(2), 321–391. <https://doi.org/10.1111/lang.12286>
- Morning, A. (2015). Ethnic classification in global perspective: A cross-national survey of the 2000 census round. In P. Simon, V. Piché, & A. Gagnon (Eds.), *Social statistics and ethnic diversity* (pp. 17–37). Springer. https://doi.org/10.1007/978-3-319-20095-8_2
- Moyer, A. (2013). *Foreign accent: The phenomenon of non-native speech*. Cambridge University Press.
- Pennycook, A. (2001). *Critical applied linguistics: A critical introduction* (1st ed.). Routledge.
- Porte, G., & McManus, K. (2018). *Doing replication research in applied linguistics*. Routledge.
- Ramjattan, V. A. (2019). Racist nativist microaggressions and the professional resistance of racialized English language teachers in Toronto. *Race Ethnicity and Education*, *22*(3), 374–390. <https://doi.org/10.1080/13613324.2017.1377171>
- Reid, K. T., Trofimovich, P., O'Brien, M. G., & Tsunemoto, A. (2021). Using task practice to reduce social influences on listener evaluations of second language accent and comprehensibility. *International Journal of Listening*, 1–16. <https://doi.org/10.1080/10904018.2021.1904933>
- Reid, K. T., O'Brien, M. G., Trofimovich, P., & Bajt, A. (2020). Testing the malleability of teachers' judgments of second language speech. *Journal of Second Language Pronunciation*, *6*(2), 236–264. <https://doi.org/10.1075/jslp.19015.tay>
- Reid, K. T., Trofimovich, P., & O'Brien, M. G. (2019). Social attitudes and speech ratings: Effects of positive and negative bias on multiage listeners' judgments of second language speech. *Studies in Second Language Acquisition*, *41*, 419–442. <https://doi.org/10.1017/S0272263118000244>
- Rothman, J., & Treffers-Daller, J. (2014). A prolegomenon to the construct of the native speaker: heritage speaker bilinguals are natives too! *Applied Linguistics*, *35*, 93–98. <https://doi.org/10.1093/applin/amt049>
- Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, *33*, 511–531. <https://doi.org/10.1007/BF00973770>
- Sheppard, B. E., Elliott, N. C., & Baese-Berk, M. M. (2017). Comprehensibility and intelligibility of international student speech: Comparing perceptions of university EAP instructors and content faculty. *Journal of English for Academic Purposes*, *26*, 42–51. <https://doi.org/10.1016/j.jeap.2017.01.006>
- Smith, G. B. (2005). I want to speak like a native speaker: The case for lowering the plaintiff's burden of proof in title VII accent discrimination cases. *Ohio State Law Journal*, *66*, 231–267.
- Stern, H. H. (1983). *Fundamental concepts of language teaching*. Cambridge, UK: Oxford University Press.
- Tiv, M., Gullifer, J. W., Feng, R. Y., & Titone, D. (2020). Using network science to map what Montréal bilinguals talk about across languages and communicative contexts. *Journal of Neurolinguistics*, *56*, 100913. <https://doi.org/10.1016/j.jneuroling.2020.100913>

- Tsehaye, W., Pashkova, T., Tracy, R., & Allen, S. E. (2021). Deconstructing the native speaker: Further evidence from heritage speakers for why this horse should be dead!. *Frontiers in Psychology*, 4467. <https://doi.org/10.3389/fpsyg.2021.717352>
- Vulchanova, M., Vulchanov, V., Sorace, A., Suarez-Gomez, C., & Guijarro-Fuentes, P. (2022). The notion of the native speaker put to the test: Recent research advances. *Frontiers in Psychology*, 1432. <https://doi.org/10.3389/fpsyg.2022.875740>
- Xie, X., & Myers, E. B. (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language*, 97, 30–46. <https://doi.org/10.1016/j.jml.2017.07.005>

Cite this article: Brown, B., Tusmagambet, B., Rahming, V., Tu, C.-Y., DeSalvo, MB., and Wiener, S. (2023). Searching for the “native” speaker: A preregistered conceptual replication and extension of Reid, Trofimovich, and O’Brien (2019). *Applied Psycholinguistics* 44, 475–494. <https://doi.org/10.1017/S0142716423000127>