A critical (re-)assessment of the effect of speaker ethnicity on speech processing and evaluation

by

Noortje de Weers

M.A., Leiden University, 2013

M.A., Leiden University, 2012

B.A., (Hons.), University College Roosevelt, 2011

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Linguistics
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY

Spring 2020

Approval

Name:	**Noortje de Weers**
Degree:	**Doctor of Philosophy (Linguistics)**
Title:	**A critical (re-)assessment of the effect of speaker ethnicity on speech processing and evaluation**
Examining Committee:	**Chair:** Nancy Hedberg Professor
	Murray Munro Senior Supervisor Professor
	Tracey Derwing Supervisor Adjunct Professor
	Yue Wang Supervisor Professor
	Elina Birmingham Internal Examiner Associate Professor Department of Education
	Molly Babel External Examiner Associate Professor Department of Linguistics University of British Columbia
Date Defended/Approved:	February 7, 2020

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

In recent years, there has been a growing interest in the bidirectional relationship between speech and social processes, as increased attention is given to how speakers' physical appearance, in combination with their accent, can influence the perception of their spoken language. Two competing theoretical frameworks have been proposed to explain conflicting findings in the existing literature: supporters of the reverse linguistic stereotyping hypothesis argue that listeners' inherent racial biases against certain groups and their speakers negatively influence their speech evaluations (e.g., Rubin, 1992; Yi, Phelps, Smiljanic, & Chandrasekaran, 2013), while proponents of exemplar-based models of perception maintain that such negative judgments reflect the cognitive consequences of incongruent face–accent pairings (e.g. Babel & Russell, 2015; McGowan, 2015). Using this debate as a point of departure, this cross-cultural and cross-linguistic investigation was designed to determine whether reported effects of speaker ethnicity also extend to online processing speeds. Two response time studies (one using photographs and one using dubbed videos of Asian and White speakers of English) were conducted in Canada, while a third study using dubbed videos of Moroccan and White speakers of Dutch was conducted in the Netherlands. Additional offline dependent measures included sentence verification scores, accentedness ratings, verbal repetition accuracy, and credibility scores. Results from the three experiments showed (1) a processing cost associated with foreign-accented and non-standard speech, but (2) no effect of ethnicity on processing speeds or on the other dependent measures. These outcomes do not support the predictions of either theoretical framework, given that both presuppose an effect of speaker ethnicity on speech evaluation. The fact that the observed null findings are consistent with some previous studies highlights the potential influence of methodological choices underlying the seemingly contradictory findings in the literature. In view of this possibility, the findings are discussed in relation to the distinction between *perception* and *interpretation*. Further research will be needed to determine the true nature and magnitude of the effect of visually based social information on speech processing and evaluation.

Keywords: audio-visual speech processing; ethnicity; accents; speech evaluation; sociophonetics; racial bias

For John and Milo.

Home is wherever you are.

Acknowledgements

I am fully aware that I would not have been able to succeed in completing this project without a battalion of people who steadfastly encouraged and supported me throughout the years. The list of people who helped me in some shape or form is very long, so I would like to use this section to specifically mention a few: my supervisor Murray Munro, who wrote me countless reference letters, gave me access to expensive equipment, provided feedback on my writing, and helped me complete this monstrosity of a thesis. Tracey Derwing, whose life advice and expert guidance helped shape this dissertation. John, the best partner I could've hoped for. You must've heard the rundown of my research so many times, you could probably have written the abstract for me! My older brother Koen, for getting a haircut he definitely did not need, just to get me a voice donor. My sister Sietske and her partner Erik, who singlehandedly recruited most of my Dutch participants in Amsterdam. My younger brother Thomas, for enthusiastically suggesting his partner as a voice donor for me (thanks J!). My parents Hettie and Paul, for teaching me to be adventurous, critical, confident, and principled. Cliff, for making me feel at home in the first months after moving to Canada, and for the many talks in his office. Molly Babel, for sharing her multi-talker-babble file with me. Stefan Grondelaers, for helping me with my data collection in the Netherlands. Saya Kawase, for volunteering to record some of her Japanese students for my first study. My 'stats guys' Neil Faught, Ian Bercovitz, and Benjamin Taft, without whom this sophisticated statistical analysis would not have been possible. Finally, I owe the success of this research to those who graciously lent me their voices and faces, and of course to those who volunteered their time to participate. After all, without them I would have had no data to analyze!

This research was also financially supported by many organizations. The Association for Canada Studies in the Netherlands was the first to provide funding, followed by Dr. Hendrik Muller's Vaderlandsch Vonds, and het Prins Bernhard Cultuurfonds. These three not-for-profit organizations financially supported my proposal before I had even begun my studies, which helped convince others that my research had merit. I am also very grateful to Simon Fraser University, the Province of British Columbia, the David See-Chai Lam Centre for International Communication, Mitacs, the Dr. Tai Whan Kim endowment, and the Social Sciences and Humanities Research Council of Canada for all awarding me scholarships to support my livelihood and this research.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AO	Audio-Only
AV	Audio-Visual
ERP	Event-related potential
EGA	Ethnic Group Affiliation
HCA	Health care assistant
IAT	Implicit Association Test
MFD	Moroccan-Flavored Dutch
RLS	Reverse Linguistic Stereotyping
SD	Standard Dutch
SNR	Signal-to-noise ratio

When it comes to social-scientific claims, exciting-seeming studies only get you so far.

–Jesse Singal

Chapter 1.

Introduction

With increasing globalization, immigration has become a major international issue. Despite the many advantages of multiculturalism, prejudice and discrimination against visible minorities remain important concerns. It is therefore useful to gain insights into how people communicate with, and develop attitudes towards each other, especially when those attitudes are prejudicial. It has already been established that linguistic minorities are often viewed negatively, which can result in discrimination in housing, courts, education, employment, and the workplace (Gluszek & Dovidio, 2010; Harrison, 2014; Lippi-Green, 2012; Zhao, Ondrich, & Yinger, 2006). In recent years, linguists have shown a growing interest in the relationship between speech and social processes, as increasing attention is now given to how a speaker's *physical appearance* (as opposed to their actual accent) can influence the perception of spoken language. Past research has demonstrated that language cannot be divorced from its social context. Indeed, numerous studies have provided ample evidence that manipulating linguistic features can influence the way a person is perceived, judged, and socially categorized (see for instance Lambert, Hodgson, Gardner, & Fillenbaum, 1960). Many of these attitudinal studies have used some form of the matched-guise technique, in which participants were asked to rate several voices on perceived personal qualities without being aware that some (or all) speech samples were in fact produced by the same bilingual speaker (e.g., Bradac, Cargile, & Hallett, 2001; Lambert et al., 1960; Lambert, Anisfeld, & Yeni-Komshian, 1965).

Surprisingly, very little research has probed the inverse relationship of visually-salient speaker information affecting speech perception (D'Onofrio, 2016; Squires, 2013; Yi, Phelps, Smiljanic, & Chandrasekaran, 2013). The few studies published on this topic seem to suggest that the ethnicity of a speaker *does* influence listener judgments; for instance, in some studies, listeners reported hearing more of a "foreign" accent when shown a picture of an Asian face (Kang & Rubin, 2009; Rubin, 1992; Yi et al., 2013), or had more difficulty understanding the speakers (Babel & Russell, 2015), even when they were actually listening to a native English speaker.

Three competing theoretical accounts have been proposed to explain these observations; while proponents of the reverse linguistic stereotyping hypothesis have interpreted this 'imagined foreign accent' as an indication of listeners' biases against Asian speakers (e.g., Rubin, 1992; Rubin & Smith, 1990), advocates of the exemplar-based model of speech perception posit that these findings actually reflect a mismatch between the listeners' expectations and the actual speech signal (e.g., Babel & Russell, 2015; Gnevsheva, 2018; McGowan, 2015). Yet another school of thought originating from psychology questions the previously-mentioned perceptual studies' findings altogether, suggesting that most of the findings are simply the result of research designs that actually measured listeners' *interpretation* of the speech signal, rather than their true perception (Firestone & Scholl, 2016; Zheng & Samuel, 2017).

Objectives and research question. This dissertation addresses this debate and its limited body of research by expanding on previous research to test the competing predictions of the exemplar theory against the reverse linguistic stereotyping hypothesis. This is done by adding visual speaker information to various differently-accented voices to investigate whether specific face–voice combinations elicit differences in (1) reaction times, (2) accentedness ratings, (3) intelligibility, and (4) credibility. This dissertation simultaneously aims to replicate the finding of Munro and Derwing (1995) that non-native voices elicit longer response times than native voices.

To investigate the effect of visual speaker information on speech perception, Chapter 2 begins by critically examining some of the main studies that have been associated with various theoretical frameworks within social psychology and sociophonetics that are of relevance to this dissertation. Chapters 3, 4, and 5 each present one of the three reaction time experiments designed to investigate the effects of stereotypes, accent, and visual speaker ethnicity on language processing and evaluation. All three experiments included a response time task and an accentedness rating task, with Experiment 2 containing an intelligibility measure, while Experiment 3 tested perceived guise credibility. The final chapter will discuss the theoretical implications of each study's findings, which will be positioned within the context of the described theoretical frameworks.

The three measures of speech that will be relevant to this discussion are intelligibility, comprehensibility, and accentedness (Kennedy & Trofimovich, 2008; Munro

& Derwing, 1995). These three reliable and valid constructs have been widely used in applied linguistics to evaluate speech. ***Comprehensibility***–also known as "processing fluency" among social psychologists–is most commonly defined as a measure of perceived cognitive effort on the listener's part, i.e., 'how difficult is it to understand this person?' and has been measured both through raters' judgments and response times. ***Accentedness*** defines how heavily foreign-accented the speech is perceived to be and is often measured through Likert scales, while the third dimension of ***intelligibility*** is often measured through recall accuracy, cloze tests, or transcription tasks, as it is meant to capture how much of the speech signal is actually understood. It should be noted that the term 'comprehension' is often used synonymously with intelligibility, but that it is not to be confused with comprehensibility.

Chapter 2.

Visual information and speech processing

Linguistic studies investigating speech perception have largely focused on auditory stimuli, while focusing far less on the role of visual cues in speech perception (Campbell-Kibler & McCullough, 2015; Cohen & Massaro, 1993). This is surprising, considering that evidence for the strong modifying influence of visual cues on speech perception was put squarely on the map as early as 1976 by McGurk and MacDonald. Their discovery, which was later dubbed the 'McGurk Effect' showed that incongruent articulatory and auditory cues alter the way auditory stimuli are perceived; when listeners were shown a face articulating /ga/ while at the same time hearing a voice producing the conflicting sound /ba/, some reported hearing a third, intermediate sound: /da/ (McGurk & MacDonald, 1976).

At this point it has been well-documented that seeing a speaker's articulatory movements (often referred to as 'visual speech') plays an important role in the speech perception process (Erber, 1975; Grant & Seitz, 2000; Peelle & Sommers, 2015; Zheng & Samuel, 2019). This is especially the case in situations where the speech signal is degraded–for example, due a noisy environment or a hearing impairment–but Audio-Visual integration has been found to contribute to the speech perception process in noise-free conditions as well (Banks, Gowen, Munro, & Adank, 2015; Cooke, Barker, Cunningham, & Shao, 2006; Sams et al., 1991).

Yet visual information involves much more than just seeing a speaker's articulators; paralinguistic or *indexical* information about the speaker can also provide the listener with valuable information on what social group the speaker identifies with, their social background, gender, and even their emotional state (Babel & Russell, 2015; Bent & Holt, 2017; Foulkes & Hay, 2015). It is important to note that the interpretation of these socially informative cues is entirely up to the listener. Whether a listener's judgment about a speaker is correct is irrelevant; it is the listener's *assumed* speaker characteristics and the listener's attitudes towards such characteristics that influence speech evaluation (Lev-Ari & Peperkamp, 2016; Niedzielski, 1999; Rubin, Ainsworth, Cho, Turk, & Winn, 1999; Rubin & Smith, 1990). The fact that these evaluations are dependent on listeners'

"cognitive representations of social constructs" was shown by D'Onofrio (2019, p. 1), who demonstrated that showing participants the same person with different clothing, facial expressions, and hairstyles could elicit different expectations about that person's (non-)nativeness within the same listener.

2.1. Speaker information through labeling

There is a growing body of research whose findings indicate that believed social characteristics of a speaker–regardless of whether they are correct–affect the evaluation and possibly even perception of speech (D'Onofrio, 2015, 2019; Hanulíková, 2018; Hay, Nolan, & Drager, 2006; Hay, Warren, & Drager, 2006; Hu & Lindemann, 2009; Johnson, Strand, & D'Imperio, 1999; Kleinschmidt, Weatherholtz, & Jaeger, 2018; Koops, Gentry, & Pantos, 2008; Niedzielski, 1999; Staum Casasanto, 2008; Strand, 1999). These findings strongly suggest that social information is indexed alongside linguistic information, and that learned associations between linguistic features and social groups influence the processing and interpretation of speech.

For most speech perception studies, the most commonly used method to provide non-visual social properties of a speaker (such as their purported gender, nationality, ethnicity, or first language) has been to provide that information explicitly to the listener. There is, however, one important caveat to this method: using generic macro-level social categories such as 'Asian,' 'old,' or 'American' to describe a speaker might merely evoke listeners' overt or stereotypical opinions of those groups as a whole, rather than tell us anything about how participants view and react to *individual* members of that group in a real-life situation (D'Onofrio, 2016; Smith & Zárate, 1992; Yook & Lindemann, 2013).

One of the more well-known examples of a study that made use of such a labeling technique was Niedzielski (1999), who argued that reported speaker information can influence how listeners perceive speech. She showed that Detroiters identified vowels differently depending on whether they thought they were rating an American or Canadian speaker. Even though the audio files had all been recordings of a fellow Detroiter, the participants were more likely to indicate hearing stereotypically 'Canadian' raised vowels when they were told that they were listening to a Canadian than when they thought the speaker was American.

Another study demonstrating the effect of learned social categories on speech categorization was conducted in New Zealand (Hay, Nolan, & Drager, 2006). The researchers primed the listeners by giving them an answer sheet headed with either "Australian" or "New Zealander," and found that even though all but one participant had in fact recognized the speaker's accent as being from New Zealand, especially female listeners were influenced by the label that they had been exposed to. That is, they were more likely to report 'hearing' more Australian-sounding tokens of synthesized vowels when the answer sheet was headed with the label "Australian" than when they were given an answer sheet headed with the label "New Zealander." This seems to suggest that listeners do not actually need to believe that the speaker is from Australia for their speech evaluation to be affected[1]; a proposition corroborated by Hay and Drager's follow-up (2010) study, which revealed that the mere evocation of a concept of a country (by strategically placing a stuffed koala or kiwi in the room) was enough to shift speakers' vowel identifications. Interestingly, however, when the same cultural primes (i.e., stuffed toys of koalas and kiwis) were used in an Australian context, this kind of perceptual bias was not observed (M. Walker, Szakay, & Cox, 2019).

Manipulating purported speaker nationality and educational level by means of short descriptive texts has also been found to influence a speaker's perceived language abilities, and even believed teaching competence (Brown, 1992). Believed nationality has furthermore been reported to influence consonant identification and intelligibility. When a sample of an American English speaker was played to fifty-three native speakers of Cantonese, the participants reported hearing instances of 'perfect pronunciation' (which they considered to be fully released stops) when they believed the speaker to be American, whereas participants who had been told that the speaker was Cantonese tended to report hearing the actual productions, which were in fact rarely aspirated (Hu & Lindemann, 2009). Another study that similarly manipulated speaker nationality had some of its native German listeners believe that the Arabic-accented German voice donor was Syrian, while others were told he was Portuguese or of an unspecified nationality (Fiedler, Keller, & Hanulíková, 2019). The researchers noted that transcription accuracy was significantly worse for the listeners who believed the speaker to be Syrian compared to

[1] Note that I do not use the word *perception* here, as the study provided no evidence of perceptual changes per se. More on this later.

the other two groups, thereby demonstrating that the listeners' linguistic expectations appeared to be informed by the speaker's believed nationality.

Besides macro-level social labels affecting listeners' judgment of speech, D'Onofrio (2015) found that presenting listeners with the stereotyped persona 'Valley Girl' or the social regional category 'Californian' affected early automatic processing as well. Her eye-tracking study revealed that both types of social information created specific linguistic expectations in the listeners, which were reflected in both their linguistic categorization of ambiguous synthesized words on a continuum, and the amount of time listeners spent looking at competing words. Even a person's name has been found to cause social characteristics to be attributed to a speaker, which in turn affects listeners' social and linguistic expectations. Names can for instance be construed as evidence of a person's ethnicity, gender (Bertrand & Mullainathan, 2004; Bursell, 2007; Riach & Rich, 2002; Senior, Hui, & Babel, 2018), and even socio-economic status (Fryer & Levitt, 2004; Staum Casasanto, Grondelaers, & van Hout, 2015).

Information about the gender of the speaker can affect speech interpretation as well. We know that listeners are sensitive to the physiological and acoustic differences between men and women, and that they store and leverage this social information to improve the processing and predicting of speech. Evidence for this has been found in infants who demonstrated the ability to discriminate between male and female voices at seven months of age (Miller, Younger, & Morse, 1982). Vowel identification has also been found to improve when the speaker's gender is correctly identified (Eklund & Traunmüller, 1997). A speaker's perceived gender can furthermore change the way listeners categorize ambiguous sounds; several studies have found that participants displayed different perceptual boundaries for /s/ and /ʃ/ depending on the attributed gender (e.g., Bouavichith, Calloway, Craft, & Hildebrandt, 2019; Strand, 1999; Strand & Johnson, 1996) or even believed sexual orientation of the speaker (Munson, Jefferson, & McDonald, 2006). Showing videos of a male and female speaker alongside ambiguous items on a vowel continuum can also shift phoneme boundaries (Glidden & Assmann, 2004), as well as merely *imagining* a male or female speaker (Johnson et al., 1999). These findings raise the question of whether they reflect a true perceptual shift, or merely a shift in the interpretation of the speech signal–a question that will be returned to in section 2.3.

2.2. Visual speaker information

As was mentioned previously, an alternative way to provide speaker information that is not as reliant on listeners' imagination is to show a photograph or video of the purported speaker along with the audio. Surprisingly, only a limited number of empirical and experimental studies explore this effect of visual speaker information on speech evaluation (D'Onofrio, 2016). As a result of this relative lack of data, considerable uncertainty remains about what processes underlie the observations that have been made so far, and how they can be explained. The remainder of this section will describe several seminal studies that have measured the effect of visual speaker information on speech evaluation according to the theoretical frameworks they ascribe to.

2.2.1. Reverse linguistic stereotyping

Most of the older studies investigating the effect of visual speaker information on the perception of speech have explained their findings though the bias model, which posits that listeners' biases and language ideologies negatively influence their evaluations of speech. Two advocates for the bias hypothesis are Rubin (1992) and Lippi-Green (2012). Rubin has argued that the findings of his various matched-guise studies could only be explained by listeners' inherent racial biases towards certain groups and their speakers. Lippi-Green (2012) and Lindemann (2002) endorse this view by positing that miscommunication with a non-native speaker has more to do with the listeners' unwillingness to put in the effort to understand the speaker than the non-native speaker's actual level of intelligibility.

The bias hypothesis was eventually further developed and renamed *reverse linguistic stereotyping* (Kang & Rubin, 2009, 2014), since the concept of linguistic stereotyping was already well-known. Linguistic stereotyping refers to listeners' linguistic stereotypes affecting their social evaluations of speakers. *Reverse* linguistic stereotyping, on the other hand, describes the phenomenon of listeners' biases about certain social identities affecting their evaluations of speech. Put differently, reverse linguistic stereotyping (RLS) is triggered by non-linguistic factors such as a speaker's ethnicity, whereas linguistic stereotyping–as the name suggests–is purely based on a person's speech patterns (Campbell-Kibler, 2009; Kang & Rubin, 2009).

In one of the most frequently-cited studies reportedly demonstrating RLS, Rubin (1992) conducted a matched-guise study at a “large southeastern [American] University” to investigate the influence of visual speaker ethnicity on teaching competence ratings, perceived accentedness, and level of comprehension (p. 514). He used a short lecture read aloud by a native speaker of English and manipulated the speaker’s ethnic identity by showing half of the participants a picture of a female East Asian lecturer simultaneously with the audio, while the other half was shown a picture of a female Caucasian lecturer. Comprehension scores were measured by means of a cloze test, in which listeners “were presented with a written text of the lecture with every seventh word deleted...*Only exact recall was scored as correct* [emphasis added]” (p. 515).

A problem with this type of comprehension measure is that it is unknown to what extent the scores are truly reflective of listeners’ understanding, and how much of the differences are simply due to the participants’ different recall abilities–independent of guise. How well listeners were able to recall target words from a four-minute long lecture tells us more about their working memory capabilities than about the degree to which they understood what had been said. Therefore, Rubin’s (1992) claim that the Asian guise undermined listening comprehension is tenuous.

Rubin (1992) further reported that the Asian guise received lower ratings on perceived teaching effectiveness, and that participants even reported hearing a non-existent foreign accent. Once again, this was a between-subjects design where listeners provided a single accentedness and teaching effectiveness rating based on a single tape recording, which meant that the listeners had no frame of reference when completing this task. This may in turn have led the listeners to make judgments according to what they thought the experimenter was looking for, i.e., their ratings were not based on their perception, but on post-hoc reflection.

Nonetheless, these results led Rubin (1992) to conclude that it was the listeners’ biases against Asian individuals that caused these distorted evaluations of speech and teaching competence. Two years prior he had drawn a similar conclusion, stating that accentedness scores (which had been obtained from single ratings as well) were negatively correlated with perceived teaching abilities (Rubin & Smith, 1990). A related observation was made in a study conducted at a major Danish business school, which had also found that the perceived English language proficiency of a lecturer did appear to

be a significant predictor of perceived teaching competence (Jensen, Denver, Mees, & Werther, 2013).

Rubin and Smith (1990) further noted that listeners who had taken more courses taught by non-native speaking international teaching assistants had better comprehension scores. This positive relationship between experience with non-native speech and comprehension has sometimes been noted by other researchers, who found a correlation between listeners' familiarity with an accent and their intelligibility scores (Derwing & Munro, 1997; Gass & Varonis, 1984; Kennedy & Trofimovich, 2008), but the evidence is weak and occasionally contradictory (Munro, Derwing, & Morton, 2006). Exposure to specific non-native accents was also reported to positively affect pronunciation ratings, as listeners with extensive experience with a given accent assigned more favorable pronunciation ratings to that accent than those who were less familiar with that accent (Carey, Mannell, & Dunn, 2011; Winke, Gass, & Myford, 2013).

In order to add generalizability to these findings, Rubin, Ainsworth, Cho, Turk, and Winn (1999) replicated the 1992 study almost a decade later, but opted to use a voice with a slight Dutch accent instead. This time, additional information about the purported speakers was provided both verbally and visually; half of the participants were shown a photograph of 'Wenshu Li,' a fictional Chinese male instructor from Taiwan, while the other half were shown a photograph of a fictional Caucasian male from Portland who was called 'Robert Wilson' (p. 5). Results showed that ascribing a Chinese nationality to the voice negatively impacted not only listening comprehension scores (once again measured by means of a cloze test), but also perceived friendliness, teaching competence, and the instructor's status. No accentedness ratings were collected. According to Rubin et al. (1999), this study thus provided additional support for the suggestion that negative stereotypes may in fact negatively impact people's judgments of international instructors.

Finally, a more recent matched-guise study focusing on health care assistants (HCAs) also concluded that listeners' negative stereotypes affected their judgments of the HCA's character and language abilities (Rubin, Coles, & Barnett, 2016). For this study, a Caucasian and a Mexican health care assistant guise were created through fake dossiers containing social information about the speakers such as their name, birthplace, ethnicity and a photograph. Listeners saw only one of the two guises. Although all had heard the same American voice, the Caucasian guise's character was evaluated more positively,

and was believed to have more of an American accent than the Mexican guise. It is important to note that this measure was once again made up of a single 'accentedness' scale item, which thus disallowed the listeners the opportunity to 'calibrate' their rating scale.

Interestingly, Rubin did not consistently find confirmatory evidence of RLS in all his studies. In a study that has thus far only been alluded to in conference proceedings (Rubin, 2012), he manipulated the purported nationality (International or American) and race (White, Asian, or Black) of a standard American-speaking guise to once again test whether the level of comprehension and degree of perceived accentedness differed as a function of either variable. Contrary to what had been hypothesized, however, race did *not* explain any of the observed variance in this study in relation to other accented speakers.

Within-subjects designs

The studies described thus far all used a between-subjects design, where each listener saw only one face and heard only one voice paired with that face. While this approach has the benefit of being less susceptible to response bias, its generalizability is doubtful, as a single speaker cannot be assumed to fully represent an accent or ethnic group.

A different approach was taken by Kang and Rubin (2009), who used a repeated measures design. Listeners heard two *different* segments of the same lecture that had been recorded by a single native speaker of English but were led to believe through fake dossiers and a photograph of an Asian and a White man that they were listening to two different teaching assistants. A noticeably accented recording was played in-between the two guises to serve as a distractor. For this study, both native and non-native speakers of English were recruited as participants, whereas Rubin's previous studies had all used native English-speaking undergraduate students. The task remained the same, with participants completing a cloze test, an accentedness rating task, and an assessment of the guises' teaching ability (through six semantic differential items such as "effective teacher or ineffective teacher" and "qualified or unqualified" (p. 448).

Kang and Rubin's (2009) main aim was to quantify their listeners' "propensity to reverse linguistic stereotyping" (RLS), i.e., they meant to measure the degree to which listeners' stereotypical views about other groups affected their evaluation of speech

produced by the two guises. Kang and Rubin (2009) made use of a series of semantic differential items that they had adapted from Zahn and Hopper's (1985) Speech Evaluation Instrument to quantity their listener's RLS scores on the *Superiority* and *Attractiveness* dimensions. To generate these scores, they subtracted the evaluations given to the Asian guise from those given to the White guise on each dimension.

Their analysis showed that listeners who had displayed reverse linguistic stereotyping on the Attractiveness dimension (e.g., they had rated the Asian guise speakers as less honest, or unkind) had worse comprehension scores for the supposedly 'non-native' Asian teaching assistant, while listeners who had shown RLS on the Superiority dimension judged the Asian guise's accent to be stronger. Finally, listeners who had demonstrated reverse linguistic stereotyping on *both* dimensions rated the Asian guise's purported 'teaching quality' much worse than the White guise, even though both guises had uttered a short section taken from the same lecture. Another notable finding was that the non-native listeners only performed worse on the cloze test than the native listeners when shown the Asian guise. Since the Caucasian guise elicited none of these effects, Kang and Rubin (2009) concluded that the effects of reverse linguistic stereotyping only seemed to apply to guises that appear to be a non-native speaker–which in their study meant a non-white individual.

A replication study using photographs of a White and a Moroccan speaker was carried out in the Netherlands by Hanulíková (2018). Using the same within-subjects design as Kang and Rubin (2009), native speakers of Dutch listened to two slightly different versions of the same native Dutch speaker reading a short story. During one story they were shown a photograph of a White male, while for the other they were looking at a (rather angry-looking) Moroccan male. Listeners completed a cloze test to measure the intelligibility of the two guises, along with scalar ratings of the purported speakers' comprehensibility and level of accentedness. Hanulíková had added listening condition (clear or embedded in speech-shaped noise) as a between-subjects variable. Contrary to Kang and Rubin's (2009) findings, Hanulíková found that neither cloze test performance nor comprehensibility ratings were affected by the speakers' ethnicity. She did, however, observe that cloze test performance was significantly worse in the noise-embedded condition, which was in line with previous research reporting the additional processing cost of noise-embedded speech. To explain her findings in light of those reported by Rubin (1992) and Kang and Rubin (2009), Hanulíková (2018) suggested that her Dutch

participants may have had more experience listening to and speaking in a non-native language than the participants from North America, which may have caused them to be less influenced by visual speaker ethnicity.

While Hanulíková (2018) did not find that accentedness ratings were affected by visual speaker ethnicity in the noise-free listening condition, those in the adverse listening condition did rate the Moroccan guise as more accented than the White guise, even though the speaker had been the same for both guises. A possible explanation for this difference between the two conditions could be that the listeners in the adverse listening condition may have misattributed their experienced speech processing difficulty to higher degrees of accentedness. That misattribution, in combination with the social expectation that the Moroccan guise is more likely to have an accent could have caused the listeners to assign higher accentedness ratings to the Moroccan guise when comparing it to the White guise, as the rating task took place after they had been exposed to both guises.

2.2.2. The exemplar theory

Although some of the findings discussed so far seem to provide support for the reverse linguistic stereotyping hypothesis, an alternative account may be offered using the exemplar theory framework (Foulkes & Hay, 2015). Studies that position their findings within this framework argue that listeners are influenced by socially meaningful information and previous experience when interpreting incoming speech (e.g., Babel & Russell, 2015; Hay, Warren, et al., 2006). Before expanding on the framework further, it is important to emphasize that there is not one single, universally accepted version of the exemplar theory within linguistics (Hay & Bresnan, 2006). There are still considerable disagreements between different versions of exemplar-based theories in terms of how information is stored in the mind and how this information is accessed (Drager & Kirtley, 2016; Guy, 2011; Johnson, 2007; Magnuson & Nusbaum, 2007).

Broadly, exemplar-based models of speech perception posit that memory traces of experienced instances of language use–including the social context–are stored in the mind in a continuously updated database of experienced tokens. These memorized representations are referred to as ‘episodic traces’ or ‘exemplars’ (Foulkes & Hay, 2015; Goldinger, 1998). During speech perception, all facets of the input are compared against this database of previously-experienced exemplars that have all been grouped into

categories (‘exemplar clouds’) to find those exemplars most similar to it. The stored tokens that most closely resemble the input are subsequently activated (Foulkes & Docherty, 2006; Hay & Drager, 2010; Rácz, 2013). Because no two individuals will have the exact same set of experiences to generalize from, each person’s cognitive representation of language will thus be slightly different (Foulkes & Hay, 2015).

Exemplars that comprise many tokens will be stronger (and thus more readily accessed) than those that were built up from fewer tokens. Additionally, the more frequently an exemplar is activated through experienced tokens, the more easily it will be activated. Since those stronger exemplars “reach full activation the fastest” [they] …therefore bias perception the most” (Drager & Kirtley, 2016, p. 6). Lower-frequency variants, on the other hand, are less robust and can eventually degrade and be forgotten due to a lack of reinforcement. This is not to say that *all* experiences are committed to memory; the relevance and importance of the information contained in the experienced tokens (i.e., social weighting) plays an important role in this selection process as well. After all, if “raw frequency of exposure” could alone determine exemplar strength, function words would have far more detailed exemplars than content words, even though they do not carry nearly as much useful information as content words (Foulkes & Hay, 2015, p. 399; Sumner & Kataoka, 2013). This dissertation aligns itself with those exemplar-based models that also allow for the encoding of more generalized, learned knowledge, as strictly episodic models (i.e., based on experiential memories) cannot explain how we are able to perceive and produce words without having encountered them before (Drager & Kirtley, 2016). Specifically, Drager and Kirtley’s (2016) account of the exemplar model has been used as the main source document for this section.

The activated exemplars also do not necessarily have to be purely linguistic; indexical information[2] can be stored alongside the linguistic token as well to create a more holistic trace (Drager & Kirtley, 2016; Foulkes, 2010; Johnson, 2006; Sumner, Kim, King, & McGowan, 2014). The encoding of indexical information into exemplars is frequently done subconsciously, which can affect the saliency of this information (Foulkes, 2010). That is to say, perceivers do not have to be consciously aware of any created associations between a linguistic item and certain social information to have their judgment affected by

[2] Note that meaning attributed to indexical information can differ from person to person, as it is personal experiences that inform indexicality (Johnstone & Kiesling, 2008).

it (Drager & Kirtley, 2016). Recent research has shown that indexical information, previous exposure, context, and whatever other information can be extracted from the speech stream appears to influence or 'bias' listeners' evaluation of a spoken stimulus (e.g., Babel & Russell, 2015; Hay, Warren, et al., 2006; Lev-Ari, Ho, & Keysar, 2018; McGowan, 2015). A good example of this approach to speech processing was given by Cai et al. (2017), who demonstrated that British and American listeners used accent information to help them retrieve the most likely meanings of words in the two dialects. Their listeners were faster and more likely to retrieve a word meaning that was congruent with the dialect in which the word had been spoken. For instance, both American and British listeners would interpret the word 'bonnet' as 'hat' if it had been spoken in an American accent, but as 'hood' (of a car) when it had been said with a British accent.

Another study noted that listeners appeared to retain episodic traces of a speaker's ambiguous utterances, as they were able to make use of that information to improve their understanding of that speaker's speech once they had been provided with the cause for the speaker's ambiguous pronunciations, e.g., speaking with a pen in the mouth (Liu & Jaeger, 2018). Listeners also seem to form expectations about sentence content based on a speaker's perceived social information, with evidence from event-related potentials (ERPs) showing that pragmatically unexpected utterances (e.g., hearing a man say "I am pregnant," or a child say "I got married last year") causes a surprise effect (Lattner & Friederici, 2003; van Berkum, van den Brink, Tesink, Kos, & Hagoort, 2008). Listeners have even been found to use information about a speaker's perceived social class[3] to predict the likelihood of that speaker using a standard or non-standard variant (Squires, 2013; Staum Casasanto et al., 2015).

A number of researchers have reported that listeners appear to make use of metalinguistic knowledge about particular social groups to predict speech patterns (Drager, 2005; Koops et al., 2008; Staum Casasanto, 2008; Strand, 1999). An example comes from a well-known eye-tracking study carried out in Houston (Koops et al., 2008). The researchers found that Houstonian listeners' expectations about front lax vowels depended on the supposed age of the speaker. A characteristic trait of the Houstonian dialect is the lack of phonemic contrast between the front lax vowels /ɪ/ and /ɛ/ before the

[3] Operationalized through "professional v. non-professional dress and middle-class v. working-class homes/backdrops" (Squires, 2013, p. 210) and "names, occupations, portraits, cars, and workplaces" (Staum Casasanto et al., 2015, p. 166).

nasals /m/ and /n/. This PIN/PEN merger is a change in progress, as this feature is now predominantly found in older Houstonians while the younger generations have started to 'unmerge' the two vowels. The eye-tracking study demonstrated that Houstonian listeners were aware of these social distributions of the PIN/PEN merger, as they made use of this information when making predictions about the speech of older and younger guises. Listeners were more likely to assume a merged system when they were shown a photograph of an old speaker than when they were shown a middle-aged speaker (Koops et al., 2008). Similar observations of age-informed social expectations were made in New Zealand English. Due to a chain shift in progress, the younger speakers are far more likely to produce raised variants of short front vowels than their older counterparts. In this context, it was also found that listeners' vowel boundaries very much depended on what they believed the age of the speaker to be (Drager, 2011; Hay, Warren, et al., 2006).

Based on day-to-day encounters with individual speakers, listeners generalize and infer information about the social and linguistic categories that these speakers belong to. As Drager and Kirtley (2016) explain, these generalized mental representations: "…contain numerous types of information about the category; including stereotypes and beliefs about the category, values associated with the category, one's past interactions with people associated with the category, and specific category exemplars" (p. 2). These learned relationships between visual and acoustic information in turn influence speech processing and evaluation. Since all this information is retrieved upon activation of the category, the listener will have formed preconceived notions about a speaker upon seeing them. These listener expectations can improve speech processing; there is a processing benefit when the speech signals match the listener's expectation, but a processing cost if the input violates the listener's expectation. This was shown by Pourtois et al. (2002), who found that listeners had more difficulty processing spoken words when the speaker's tone of voice and facial expression did not match. ERP data provided by Hansen et al. (2017) similarly showed that unexpected face–voice pairings were related to "more effortful cognitive processing" (p. 513) than expected pairings.

The practical use of this 'database' of social and linguistic knowledge was further demonstrated in a study by Staum Casasanto (2008), who found that her listeners used linguistic knowledge about two specific ethnic groups to determine whether ambiguous words contained t/d deletion or not. In one of the experiments, listeners were tasked with deciding as quickly as possible whether a sentence they heard made sense or not (e.g.,

'*the mass/mas[t]…probably lasted through the storm'* or *'the mass/mas[t]…probably lasted an hour on Sunday*'). Staum Casasanto observed that listeners were faster at responding to a sentence where t/d deletion was necessary for the sentence to make sense when it was accompanied by the photograph of a Black face. When the speaker was White, however, listeners reacted faster to sentences where the ambiguous word did *not* require t/d deletion to make sense. These results showed that listeners associated t/d deletion more with African-American speakers than with White speakers, probably because African American Vernacular English speakers use the deleted variant more frequently than White speakers do (Labov, 2001; Staum Casasanto, 2010).

When there is a lack of exposure to members of a certain group, listeners may fall back on encoded stereotypical or imitated features of that group to facilitate processing. This was demonstrated by Neuhauser and Simpson (2007), who discovered that German monolingual listeners were better at recognizing and naming *imitations* of American and French accents than the authentic accents (with which they had far less experience). Similarly, another study showed that listeners who had less experience with Chinese-accented speech were worse at distinguishing between an imitated Chinese accent and an authentic one than more experienced listeners (McGowan, 2016).

2.2.3. Bias and stereotypical expectations

According to the exemplar theory, listeners hearing a non-native accent from an Asian language (either in person or through broadcasting media) will automatically activate the generalized association of 'Asian = accented' in their mind when seeing an Asian face. As Squires (2013) explains: "in a language processing situation, a social cue [such as speaker's ethnicity] can shift one's perception toward the linguistic variant that has been experienced as more frequently linked with that social property" (p. 202). While this association could still be labeled as stereotypical, it is in this case based on the listeners' frequency of exposure and their subsequently stored tokens. This facilitating effect of frequently-experienced exemplars is demonstrated by Walker and Hay (2011) and Kim (2016), whose studies both observed that words that are more often said by younger speakers were recognized significantly faster and more accurately if they were said by younger-sounding voices compared to older ones, while words that are more often said by older speakers were in turn more easily processed when said by older speakers.

Things become more interesting when listeners continue to have certain expectations *despite* having experienced many interactions that showed the contrary; this suggests that something other than frequency is at play. As Bent and Holt (2017) posit, "the extent to which social information is encoded in memory may depend not just on pure quantity of experience, but also upon the quality of the experience and the attentional resources allotted to the social-indexical category" (p. 2). Thus, low perceived social salience, or a lack of attentional weight given to 'unexpected' experienced tokens may contribute to the listener failing to adjust expectations accordingly (Dahan, Drucker, & Scarborough, 2008). There is indeed evidence suggesting that we are better at remembering information that confirms our expectations than information that disconfirms them (Fyock & Stangor, 1994). In these cases, listeners' expectations appear to be based not only on frequency of exposure, but also on their stereotypical beliefs, which influenced their retention of specific experienced exemplars. A good example of this comes from Babel and Russell (2015), whose listeners' stereotyped expectations about ethnically Asian speakers appeared to negatively affect their processing efficiency and accentedness ratings. The study was conducted in the multi-ethnic and linguistically diverse city of Metro Vancouver, where immigrants make up 40.8% of the population (Statistics Canada, 2017c). A large proportion of recent immigrants are Chinese, but there are also many second- and third-generation immigrants with Asian heritage. This means that Vancouverites should frequently encounter both native and non-native Asian speakers of English. While older Asian speakers are less likely to be native speakers of English, interlocutors have no a priori way of knowing whether young Asians are native or non-native speakers of English. Interestingly, individuals with an Asian background are nonetheless often erroneously expected to be non-native speakers of English–which according to the bias paradigm may trigger reverse linguistic stereotyping.

Babel and Russell (2015) therefore set out to explore whether knowing the ethnic background of a speaker would influence the perceived intelligibility of that person's speech. Forty native speakers of English were asked to transcribe sentences embedded in pink noise. The researchers used twelve self-identified native speakers of English, six of whom identified as Chinese and the other six as Caucasian. Half of each speaker's sentences were accompanied by their photograph, while the other half were shown alongside three fixation crosses. Before beginning the task, participants were primed verbally, as they were "casually informed that all the speakers they would hear were from

Richmond, BC" (p. 2826), which is a municipality within Metro Vancouver known for its large (mostly Asian) immigrant population[4]. This prime was meant to activate expectations of non-nativeness in the listeners. After transcribing the sentences, the same participants also judged the twelve speakers on perceived accentedness and completed an implicit association test (IAT) and a stereotype questionnaire to measure their implicit and explicit racial attitudes.

Results showed that the Asian Canadians received higher accentedness ratings than the White Canadians, and that this difference grew in the Audio-Visual condition, where accentedness ratings *decreased* for the White Canadians but *increased* for the Asian Canadians, even though all speakers were native speakers of English. The authors also observed that the intelligibility of *only* the Chinese-Canadian voices dropped in the AV condition, while there had been practically no difference in transcription accuracy between the White and the Chinese Canadian voices in the Audio-Only condition. Curiously, the participants who reported having a predominantly Asian Canadian social network were also the ones who had more difficulty transcribing the Asian Canadian speakers in the AV condition (compared to those who had described their social networks as mostly White Canadian).

Since listeners' biases about Chinese Canadians (as recorded through the IAT[5] and stereotype questionnaire) were not found to predict their performance, Babel and Russell (2015) argued that the lower intelligibility of the Asian Canadians could not simply be attributed to a lack of effort on the listeners' part. Instead, they ascribed these intelligibility differences to the listeners' stereotyped associations, which were activated when they were shown the photographs. They posited that the Chinese Canadians' visual primes activated stereotypes in the listeners' minds that led them to anticipate a non-native accent. The subsequent misalignment between this anticipated non-native accent and the actual observed native speech caused the decreased level of intelligibility for the Asian

[4] According to the 2016 Census, more than half of Richmond's population (60.2%) has immigrant status, and 56% of those immigrants were born in China. Of those living in Richmond, 64.1% are first generation (i.e., born outside Canada), 22.4% were born in Canada with at least one parent born outside Canada, and 13.6% were born in Canada with both parents born in Canada (Statistics Canada, 2017b).

[5] IAT scores from a single sitting should not be used to divide participants into 'more biased listeners' and 'less biased listeners' (see the critical evaluation of Yi et al. (2013) at the end of this section for a more detailed explanation as to why).

Canadian speakers. Since there was no such mismatch between expectation and speech signal for the White Canadians, their transcription scores did not suffer in the Audio-Visual condition. This mismatch effect could also be used to explain the higher accentedness ratings.

A few years later, Babel and Mellesmoen (2019) carried out a replication study in Vancouver. This time they used videos of two Asian and two White individuals who spoke English with either a native or non-native accent. Once again, the four speakers' voices were embedded in pink noise to test the effect of visual speaker ethnicity on the transcription accuracy of high and low predictability sentences in adverse listening conditions. Findings partially supported an exemplar-based account, as the intelligibility of the stereotypically 'unexpected speakers' in the Vancouver context (i.e., the White speaker with a foreign accent and the Asian speaker with a native accent) did indeed decrease in the low predictability condition compared to sentences with high predictability. Yet while an exemplar theory of speech perception also predicted improved transcription for congruent face–voice combinations, no such facilitative effect was observed.

This stereotypical expectation to hear foreign-accented speech when seeing an Asian face is not unique to listeners in the Vancouver context. An influential psychology study conducted in the US revealed that Americans also still consider the prototypical American to be White, even though the United States is a multicultural society. While Asian and White Americans were rated as 'equally American' on a conscious level, implicit association tasks still showed that Americans were faster at associating 'Asian' items with 'foreign' items (Devos & Banaji, 2005).

This 'American = English speaker' association is also not unique to adults; it has also been observed in five-year-old children who were living in the greater Chicago area. When asked to categorize various guises as either American or Korean (based on a voice and a photograph), both American and Korean children were found to disregard visual speaker ethnicity and instead based their decision on language use. Any guise that spoke in an unfamiliar language or accent–i.e., French, Korean, or Korean-accented English–was rated as Korean (DeJesus, Hwang, Dautel, & Kinzler, 2018). However, some of the older nine-year-old participants had in fact started to take visual speaker ethnicity into account as well when deciding on the nationality of the guises, suggesting that associations between language, ethnicity and nationality are learned very early on.

Several infant studies confirm this, as they have shown that babies as young as eleven months old already associate specific ethnicities with specific accents (May, Baron, & Werker, 2019; Uttley et al., 2013) and that their speech processing appears to be affected by visual speaker ethnicity as well (Weatherhead & White, 2018).

Additional support for the assertion that listeners tend to expect Caucasian speakers to be native speakers was reported by Gnevsheva (2018), who used a between-subjects design to compare the anticipated accentedness scores of White German speakers among New Zealand listeners in a Video-Only condition against the accentedness scores given to these same German speakers in an Audio-Only or Audio-Visual condition. She found that the soundless videos of the White German speakers elicited lower anticipated accentedness scores than when their actual voices were played in an Audio-Only condition. The German speakers received the highest accentedness ratings in the Audio-Visual condition, which strongly suggests that there was indeed some degree of expectation violation, i.e., the New Zealand listeners did not expect the Caucasian speakers to have a foreign accent at all, so this surprise effect resulted in higher accentedness ratings. Note that this experiment compared participants' *expected* accent ratings (based on a soundless video) against other listeners' actual evaluations of accentedness based on auditory input, which may not be a fair comparison.

If Rubin's American listeners had a similar association of 'Asian = foreign' and 'White = native', they would have expected the Asian guise to be a non-native speaker of English. The subsequent mismatch between their activated expectations and the actual speech signal may thus have been the reason for the observed higher accentedness ratings and worse comprehension scores for the non-white guises. Rather than interpreting the listeners' reactions as prejudiced against Asian individuals, we could thus also argue that they simply made incorrect stereotypical inferences about the listeners, and that it was their violated expectations that 'got in the way' of their processing efficiency. This exact suggestion was made by McGowan (2015). To test this hypothesis, he devised a between-subjects design that enabled him to compare the predictions of the exemplar model to the reverse linguistic stereotyping model. After telling his participants that they would be hearing a graduate student instructor who could be either an American or a native Mandarin speaker, McGowan had his participants transcribe sixty Mandarin-accented sentences that had been embedded in multi-talker-babble. Depending on the group they were assigned, listeners were shown a photograph of either a Caucasian face,

an Asian face, or a stylized silhouette. The latter group's ratings were meant to provide baseline transcription accuracy scores as the silhouette contained no visual cues about the speaker's social identity to influence the transcription accuracy (it is worth noting that none of Rubin's studies had such a comparison condition).

Rubin's RLS model predicts that the Asian guise would elicit lower transcription accuracy scores because of the negative stereotypes associated with Asian instructors; the participants are expected to reject their share of the communicative burden when they are given any indication that they are confronted with a non-White speaker, regardless of the actual accent. Their subsequent reduced attention would then result in lower transcription accuracy scores for the Asian guise, while no worse performance should be seen in the White face or silhouette conditions, as those two guises were not expected to evoke any negative bias.

The exemplar theory, on the other hand, predicts that the Asian face will bring about *enhanced* transcription accuracy scores, since its pairing with a Mandarin accent should be congruent with listeners' expectations. Presenting the (incongruent) Caucasian guise, on the other hand, likely violates listener expectations and will result in the activation of the wrong base levels. This incongruent pairing is therefore predicted to have an *inhibiting* effect on transcription scores (McGowan, 2015).

Results showed support for the exemplar theory, rather than for reverse linguistic stereotyping; instead of the Asian guise inducing lower transcription accuracy scores due to listeners' supposed negative biases, participants presented with the congruent condition (i.e., Mandarin-accented speech paired with the Asian face) performed significantly better on the transcription task than those presented with the incongruent Caucasian face or the silhouette. These results suggest that the listeners had stereotyped expectations about what Asian and White people 'should sound like,' as presenting them with stimulus pairings that matched these expectations facilitated their understanding of the speech.

Another interesting finding was that experience with foreign accents improved transcription accuracy. McGowan (2015) had conducted a pre-test to divide participants into those who had more or less experience with Mandarin accents. The 'experienced' participants (as determined by a self-report and performance on an accent authenticity

detection task) were significantly better at transcribing the Chinese-accented speech than the 'less experienced' participants, who had had less exposure to Chinese-accented speech, and were consequently assumed to be less adept at recognizing the difference between an authentic and imitated accent. These findings further support the exemplar theory, as more experience with accented speech results in stronger exemplars, which in turn helps with anticipating and better understanding accented speech.

Based on these results, McGowan (2015) rejected the RLS hypothesis, and instead suggested that Rubin's (1992) results could be explained within the exemplar theory's framework. He argued that both studies reported lower performance scores when the listeners were misled about the identity of the speaker through an incongruent guise, which in Rubin's study happened to be the Asian face. When the listeners were presented with a congruent guise, both studies recorded higher performance scores. McGowan (2015) therefore maintained that rather than Rubin's (1992) findings providing evidence for reverse linguistic stereotyping, they simply demonstrate that incongruent face–accent pairings result in lower intelligibility scores. McGowan (2015) is careful to point out that his study does *not* suggest that there is no bias whatsoever against non-native speakers; he simply did not find any evidence for negative bias playing a role at either the perceptual or attentional level in his experiment.

This effect of unmet expectations causing worse performance has been observed in social psychology experiments as well, where it has been explained through the 'expectancy violations theory,' which posits that "when people encounter others whose appearance and accent do not match…such violations should produce more extreme outcomes than situations matching expectations" (Hansen, Rakić, & Steffens, 2018, p. 1002). Practically, this means that when expectations are violated positively (e.g., listeners expect to hear a foreign accent based on the speaker's ethnic appearance, but do not hear one), ratings of that person turn out better than the ratings of people whose accent and appearance were congruent to begin with (Bettencourt, Dill, Greathouse, Charlton, & Mulholland, 1997; Hansen, Rakić, & Steffens, 2017; Hansen, Steffens, et al., 2017). For instance, Hansen et al. (2018) found that when they showed German listeners a photograph of a Turkish man before playing a short recording of a native German voice, the incongruent face–voice pairing negatively affected the hirability evaluations of the guise. Yet when the presentation order of this incongruent pairing was reversed and the

Turkish face was only shown *after* the recording, the positive expectancy violation caused the incongruent Turkish guises to receive the highest hirability ratings instead.

Finally, I would like to critically re-evaluate a study that has frequently been cited as evidence for the bias hypothesis. Yi, Phelps, Smiljanic, and Chandrasekaran (2013) carried out a study on native and non-native speech in noise to test the effect of visual cues on native listeners' transcription accuracy. Two White native English speakers and two native Korean speakers (one male, one female) each recorded forty sentences which were embedded in multi-talker-babble and played to twenty-one native monolingual speakers of English. The AV condition showed a video of the speakers, while the AO condition showed a fixation cross instead of the video. A within-subjects design was used, which meant that each participant saw all four speakers in both AV and AO conditions. Since there were forty target sentences in total and no sentences were repeated, this meant that each listener transcribed only 5 sentences per condition (e.g., the female native speaker in the AV condition). The listeners also completed an implicit association test (IAT) to measure the strength of their 'Asian = foreign' and 'Caucasian = American' associations. Six separate raters judged each of the four speakers' forty sentences on accentedness in both Audio-Only and Audio-Visual conditions, so an overall accentedness score could be calculated for each of the four speakers in the two conditions.

The exemplar theory posits that comprehension scores should benefit from additional visual information, provided that this visual information is congruent with the listeners' expectations. Since the speakers' ethnic appearance did indeed 'match' their accents (i.e., Korean face + Korean-accented speech; White face + American speech), the finding that the addition of the videos had a beneficial effect on the comprehension of *both* native and non-native speech was in line with this prediction. If the listeners had simply reduced their effort to comprehend the speech as soon as they saw an Asian face–as the bias hypothesis predicts–then transcription accuracy should have worsened in the Audio-Visual condition compared to the Audio-Only condition. Interestingly, the authors did note that the AV benefit was actually greater for the native voices than for the two non-native Korean voices (Yi et al., 2013)–an observation that was also made in a later study (Xie, Yi, & Chandrasekaran, 2014).

The implicit association test results added another layer to this last finding, as its results suggested that participants with stronger 'Asian = foreign' and 'White = American' associations had relatively more difficulty with non-native speech in the AV condition. The authors therefore concluded that these findings provided evidence of "non-linguistic visual bias affecting speech processing" (p. 392). However, the validity and reliability of implicit association tests has been the subject of heated debates. Several serious problems with the IAT have been identified (Fiedler, Messner, & Bluemke, 2006), and several meta-analyses have been carried out to examine its reliability in more detail. None of these meta-analyses found convincing evidence that IAT scores actually predict real-world behavior. At this point it is not clear what exactly IATs test, as several studies were able to manipulate IAT scores through simple interventions that had nothing to do with implicit bias. Singal (2017) concludes in his extensive review of the debate around the validity of the IAT that "there's no good, empirically backed reason to assume that a given IAT score reflects your actual level of implicit bias, as opposed to a noisy mishmash of other stuff." Yet more to the point, even the creators of the IAT have conceded that a single IAT test *cannot* (and therefore should not) be used to predict an individual's subconscious racial biases (Singal, 2017). Therefore, tagging participants with high IAT scores as 'more implicitly biased against Asians' than those with lower scores, and using these scores as a binary variable to divide participants into two groups may yield no real insight into how these participants would respond outside of a lab setting. Still, Yi et al.'s (2013) conclusion has been cited by multiple researchers as evidence of bias against Asians affecting speech processing.

Although the argument against using IATs has already been made, it is worth mentioning an additional criticism about the type of IAT that was used in Yi et al.'s (2013) experiment. The test consisted of photographs of ten Asian and ten White faces, together with "[p]ublic domain images of ten iconic American scenes (e.g., Grand Canyon, Statue of Liberty) and ten non-American foreign scenes (e.g., Eiffel Tower, Angkor Wat)" (p. 389). Participants in the congruent condition were asked to group the Asian faces together with the 'non-American foreign scenes,' whereas those in the incongruent condition had to group the Asian faces and typical American scenes together. This task seems straightforward, but the first example of a 'typical foreign scene' illuminates that the category 'foreign' could potentially have elicited various interpretations among the listeners. The Eiffel Tower was probably meant to evoke an image of the country France,

which is indeed a ‘foreign scene’ to an American listener. However, it may also evoke an image of a ‘typical French person,’ who is decidedly *not* Asian in appearance. Thus, asking listeners to lump Asian faces together with a foreign scene that in their mind’s eye is predominantly Caucasian is counter-intuitive, but in this IAT it was considered to be the congruent condition. Another example of a foreign scene was a photograph of pyramids. Again, the speed at which someone can relate an Asian face with a pyramid does not necessarily reveal anything about that person’s associations with and biases against Asian individuals. In sum, apart from the limitations of IAT, the particular associations Yi et al. (2013) chose to test deserve skepticism. The limited number of observations per cell (twenty-one participants for the intelligibility and IAT test, six for the accentedness ratings) also raise methodological concerns.

Yi et al.’s (2013) second finding that has been frequently cited as evidence of listener bias was that the Korean speakers received *higher* accentedness ratings in the AV condition relative to the AO condition, whereas the opposite was true for the White native English speakers–their perceived accentedness was actually *lower* in the AV condition. A closer look at the actual differences in accentedness ratings between the two conditions reveals that these differences were extremely small (only 0.9% difference for the English speakers, and 1.8% for the Korean speakers). Since the researchers did not follow up on the significant interaction with a post hoc test or reported effect sizes, these tiny differences in accentedness ratings between the AV and AO conditions should not be taken as evidence of a meaningful difference. Another important methodological decision to note here is that only six participants provided accentedness ratings for the four speakers in both the Audio-Visual and Audio-Only conditions, i.e., each participant rated 320 tokens. Although this many accentedness ratings per speaker is certainly preferable over a single Likert rating, the fact that only six individuals provided these ratings casts further doubt upon the generalizability of these findings.

Perhaps due to the methodological decisions made by Yi et al. (2013), more recent studies investigating the effect of visual speaker ethnicity on perceived accentedness failed to replicate their findings. Karpinska (2019), for instance, used a blocked design to measure whether Japanese native speakers would judge the accentedness of six native speakers of English differently as a function of their ethnicity. After establishing each listener’s baseline ratings in an Audio-Only condition, listeners were shown realistically dubbed videos of either Caucasian or Asian individuals. Both conditions featured the same

native English voices. A third control group heard these same voices without a visual cue. Contrary to what had been hypothesized, listeners did not give the incongruent Asian guises higher accentedness ratings than the congruent White guises. Similarly, McCrocklin, Blanquera and Loera (2018) failed to find an impact of visual speaker ethnicity on accentedness ratings in an experiment for which they had used native speakers of English, Spanish and Mandarin in combination with photographs of White, Hispanic and Asian faces, respectively. Additionally, congruent or incongruent name and accent pairings (e.g., 'John' speaking with an native English or Mandarin accent) were not found to affect accentedness ratings either (Senior et al., 2018). Interestingly, Rubin and Smith (1990) also found no evidence of visual speaker ethnicity differentially affecting accentedness when they exposed listeners to either a mildly or a heavily accented voice in combination with either a photograph of a White or Asian woman.

In summary, there is currently not enough convincing evidence to make any categorical statements about the effect of visual speaker ethnicity on either perceived accentedness or intelligibility, as there are serious methodological issues with some of the more frequently-cited studies in the field. The conflicting findings reported here strongly suggest that the studies' different research designs may have been partially to blame for eliciting these varying responses. As observed by various researchers (see for instance Levi, Winters, & Pisoni, 2007; Vaughn & Baese-Berk, 2019; Zheng & Samuel, 2017), accent ratings appear to be context-dependent, as they have shown to be differentially impacted by various factors such as presentation modality (photos vs. videos vs. no visuals at all), the number of items rated, and even the order in which speech samples are presented, to name but a few.

2.3. Perception vs. interpretation

A third way of interpreting the results that have been presented so far within the RLS hypothesis or exemplar-based framework has been offered by Firestone and Scholl (2016), who made the strong claim that, despite a large body of research arguing otherwise, no study had in fact provided convincing evidence of top-down effects on visual perception. Although Firestone and Scholl's (2016) paper was originally meant to critique the methodologies of *visual* perception studies, their list of six common pitfalls that they believe most top-down studies fall prey to has been acknowledged by a few speech

researchers as well (Cutler & Norris, 2016; Zheng & Samuel, 2017). The three pitfalls most pertinent to speech perception studies will be discussed below.

Perhaps the most important requirement of any perception study is the ability to discern where perception ends, and memory begins. Take for instance Niedzielski's (1999) finding that when Detroiters thought they were listening to a Canadian, they were more likely to match the vowels they had heard in a fellow Detroiter's speech to synthesized vowel tokens that contained Canadian raising. The author (and many others with her), interpreted this finding as evidence that vowel *perception* had been influenced by speaker nationality. In this experiment, listeners were tasked with remembering a vowel that had been featured in a specific target word, which had been embedded in a sentence. Listeners were able to choose the best match only between the remembered vowel and six different synthesized vowels after hearing the entire sentence. The task thus relied heavily on listeners' memory, which leaves open the question of whether it was truly perception that had been affected, or perhaps their memory (Cutler & Norris, 2016). Other, similar vowel-matching studies making use of labeling (e.g., Hay & Drager, 2010; Hay, Nolan, et al., 2006; Hay, Warren, et al., 2006; McGowan & Babel, 2019) have also claimed to provide evidence that "social information can alter and override listeners' use of phonetic detail" (McGowan & Babel, 2019, p. 14). Yet since these designs were equally reliant on memory, it remains unclear to what extent these studies are truly measuring a perceptual shift, and to what extent their findings are reflective of some kind of social evaluation on a more conscious level.

This brings us to the second pitfall. Firestone and Scholl (2016) caution that any perceptual study should "disentangle post-perceptual judgment from actual online perception" (p. 18). This problem arises when participants are asked to report on what they 'perceived' after exposure to a stimulus. Although this method has been widely used, Firestone and Scholl argue that it is problematic, as it remains unclear whether any changes in these reported post-stimulus observations are reflective of actual perceptual changes, or simply of a change in the participants' judgments. An example of such a study where the distinction between perception and interpretation is blurred would be a study in which participants have to pick the most 'attractive-sounding voice' out of three voices. Firestone and Scholl (2016) would posit that this type of design does not measure actual perception, as we cannot *perceive* attractiveness; we can only infer it. Furthermore, note that memory would play a role in this hypothetical task as well, since listeners would have

to keep their representations of all three voices activated in their minds in order to make a comparison.

Another example to illustrate the challenges of using Likert scales to get accurate information about a person's perception is eliciting 'perceived' pain levels from patients, as their answer will depend on a variety of factors such as their frame of reference (how much does it hurt compared to pain they have felt previously?), their propensity to use the extremes of a scale, and personal motivations (trying to downplay the pain to put on a brave face, or exaggerating it in the hopes that their pain will be taken more seriously). It is therefore dangerous to use these ratings as a reflection of their actual perceived level of pain. Instead, Likert scales are impressionistic measures, rather than perceptual ones. Although there is no denying that guise ethnicity has been found to affect speech and character ratings, whether these results are truly indicative of *perception* being influenced by guise ethnicity is not as clear as it is made out to be by some authors. An example of such a false equivalence of perception and interpretation can be found Rubin and Smith (1990), who noted that:

> the degree to which subjects *believed* the speech samples were accented (as opposed to the level of actual accent) was a good predictor of how they rated the NNSTA's teaching ability. The higher the *perceived* level of accentedness, the lower the teaching ratings [emphasis in original] (p. 349).

Another good example illustrating that what listeners *perceive* in a perceptual task is not necessarily the same as what they *believe* they heard comes from McGowan and Babel (2019). They found that even though the listeners' performance in an AXB vowel matching task was similar between the experiment's two guises, the listeners still indicated in a post-test interview that they believed the guises had been different (even though they were not).

A final pitfall that will be discussed here revolves around demand characteristics. This experimental artifact comes into play when participants think they know the goal of the study, which could cause them to "adjust their responses (either consciously or unconsciously) in accordance with their assumptions about the experiment's purpose" (Firestone & Scholl, 2016, p. 10). Within-subjects designs tend to be more susceptible to various types of response bias, as listeners' exposure to all conditions makes it easier for them to recognize the study's (often rather conspicuous) manipulations and to

subsequently make an educated guess about the study goals. As a result, when participants believe that a study's purpose is to tap into their biases, they may attempt to construct a better image of themselves by giving politically correct responses instead of honest ones (i.e., social desirability bias).

These types of biases are less of a problem in between-subject designs, as participants do not get to see all manipulations, and are therefore less likely to correctly guess the purpose of the study. That is of course not to say that participants in between-subjects designs do not venture any guesses. Even in a between-subjects design, showing a static image of a speaker might cause the participants to speculate that the goal of the study is related to the image presented.

Taking these three pitfalls into account, two researchers extended Firestone and Scholl's (2016) argument to speech perception as well, and similarly argued that a large number of studies that claimed to have found an effect of speaker face on speech perception really just measured listeners' *inferences* about the speech signal, rather than actual speech perception (Zheng & Samuel, 2017). They ran six related experiments specifically developed to test whether visual speaker ethnicity affected accentedness judgments.

As their first experiment, Zheng and Samuel (2017) replicated Rubin's (1992) between-subjects design. Instead of a passage, they chose to create randomized eight-step continua of the words *cancer, theater,* and *thousand*, with American English on one end, and Mandarin-accented English on the other. Listeners were then asked to rate the four middle four steps in the continuum (i.e., that were most in-between the American and Mandarin-accented voice) for all three words on a four-point accentedness rating scale. This larger number of ratings made Zheng and Samuel's accentedness measure more nuanced than Rubin's (1992), as his accentedness scores had been based on only a single semantic differential item. Nonetheless, their results were consistent with what Rubin (1992) had reported; participants who had been shown the Asian face rated that guise as more accented than those who had seen the Caucasian face.

However, when Zheng and Samuel (2017) had the same listeners also rate the other, remaining guise (thereby making it a within-subjects design), it was now the Caucasian face that received the highest accentedness ratings. The authors hypothesized

that this reversal might be due to overcompensation on the listeners' part, with the listeners actively changing their responses to avoid bias once they figured out the guise manipulation. Regardless of whether the observed differences in accentedness judgments were due to listeners shifting their judgments, these findings do show that the higher accentedness ratings were not uniquely tied to the Asian guise.

Because Zheng and Samuel (2017) hypothesized that the static images evoked different ratings simply because participants could guess what the study was about, a second, identical experiment was conducted on a different set of listeners who were shown dubbed videos instead of photographs. Findings showed that the Asian guise videos were rated as slightly more accented-sounding than the White guise in both blocks, which seems to corroborate the reverse linguistic stereotyping hypothesis.

In the third experiment, the researchers added a contrast manipulation by adding the most native-sounding step to the White face block, and the most foreign-accented step to the Asian face block. As expected, these 'anchor points' caused a classic contrast effect, with the listeners now rating the ambiguous items paired with an Asian face as less accented, and the ambiguous items paired with the White face as more accented. Yet when the researchers repeated the exact same experiment in a pseudo-randomized design where the Asian and White face were mixed, this effect of face disappeared.

In the last two experiments, Zheng and Samuel (2017) demonstrated that repeated exposure to a clearly accented token resulted in subsequent items being rated as less accented, while exposure to a native-sounding adaptor increased the accentedness ratings of the other items. Yet when they used Asian and White Faces as adaptors instead, no such contrastive effect was found, which indicated that accent perception was unaffected by visual information.

Perhaps the most significant finding of these six experiments was that study design seemed to shape listeners' accentedness responses to a large degree; Zheng and Samuel (2017) were able to replicate Rubin's findings only when they used a between-subjects design. The effect of adapters on accentedness ratings served as a reminder that accentedness ratings are not stable independent measures that are impervious to contextual influences. As such, extreme care must be taken to not bias the listeners' ratings through accidental adaptation or ordering effects.

2.4. Interim summary

The sociophonetic literature reviewed above has shown robust evidence of the significant effects that indexical speaker information can have on phoneme categorization, word recognition, and overall speech evaluation. Yet considerable uncertainty remains about what exactly is causing these observations, and whether processing is also impacted by this indexical information. Although Rubin's findings have often been cited as evidence of listener bias, the divergent findings in more recent studies have cast serious doubt on his conclusions. As was shown repeatedly, the purported effect of visual speaker ethnicity on speech perception seemed very much dependent on the study design and task types that were used. Furthermore, the majority of studies discussed so far proved to be susceptible to one or multiple pitfalls listed by Firestone and Scholl (2016), which brings the validity of the interpretations into question.

The experiments that will be described below attempted to address Firestone and Scholl's call for separating "post-perceptual judgment from actual online perception" (p. 18) by measuring response latencies in a sentence verification task. The distinction between online and offline measures has remained surprisingly vague in the speech perception literature[6], but there seems to be general agreement that online measures capture real-time processing efficiency *during* the stimulus presentation (through, for instance, ERP recordings or eye-tracking), while offline measures test comprehensibility *after* the presentation of the stimulus (Brückner & Kammer, 2017; Fernald, Perfors, & Marchman, 2006; MacWhinney, Feldman, Sacco, & Valdes-Perez, 2000; Münster & Knoeferle, 2018; Shapiro, Swinney, & Borsky, 1998; Zangl & Fernald, 2007).

Sentence verification tasks require participants to make quick judgments about a series of sentences, i.e., they must determine as quickly as possible whether a sentence they read or hear is grammatically correct, or whether a sentence is true or false. The time it takes to respond to the stimulus is measured in milliseconds. According to the traditional distinction between offline and online measures, then, a sentence verification task should

[6] For instance, although the book *The on-line study of sentence comprehension: Eyetracking, ERPs and beyond* (Carreiras & Clifton, 2004) mentions 'off-line' seventeen times contrastively with online processing, none of its authors provided a clear definition of either construct.

be characterized as an offline measure, as it measures post-sentence reaction times[7]. However, it appears that Firestone and Scholl (2017) were not using the same narrow definition of "online perception" as psycholinguists do. Rather, they appeared to use the word 'online' to put emphasis on the 'actual' perception in contrast with the 'offline' interpretation of that experience afterwards.

2.5. Measuring comprehensibility through reaction times

This brings us back to response time studies. It can be argued that sentence verification tasks fall somewhere in-between online and offline tasks. Where self-paced tasks such as accentedness ratings or cloze tests allow participants to consciously reflect on, or even change their response to a more socially appropriate one, sentence verification tasks force participants to provide an immediate and automatic response. As such, they are more likely to tap into subconscious expectations and listener prejudices without falling prey to response bias to the same extent that offline tasks do. Cloze tests and Likert scales furthermore both rely more heavily on recall compared to a verification task. An example of a memory-reliant measure in linguistics is the use of Likert scales to collect accentedness scores, which are usually collected after the listener has completed the task, and often use a single or a few recordings per speaker as prompts to remind the listener of the speaker's accent (Vaughn & Baese-Berk, 2019). Based on these few selected items, the listeners are then asked to generate an overall single accentedness score for that speaker (e.g., Babel & Russell, 2015; Hanulíková, 2018; Rubin et al., 2016).

Although transcription tasks and verification tasks both require an immediate response following a sentence, the process of writing or typing out a sentence takes longer than a single button-press that is required for a verification task. The more time that elapses between the end of the stimulus and the completion of the response, the more susceptible the response becomes to additional cognitive processes that may ultimately alter the response or evaluation. Effectively, where response times elicit listeners' subconscious or implicit attitudes[8], evaluative measures such as Likert scales are more

[7] Münster and Knoeferle (2018) and Millotte, René, Wales, and Christophe (2008) also consider reactions times to be online measures.

[8] It is worth noting that not all researchers see these two terms as synonymous. According to Pantos (2019), those in the attention research tradition "agree that implicit measures avoid conscious

reflective of explicit attitudes, as they allow listeners to introspect on their own judgments and thus elicit more controlled answers.[9]

So far, there has been little investigation into the effect of visual speaker ethnicity on comprehensibility through processing times. To the best of my knowledge, only Staum Casasanto (2008, 2010) and Squires (2013) made use of response times to measure the effect of social speaker information on speech processing. To address this gap in the literature, three thematically linked reaction time studies were created for this dissertation project to further test the effect of visual speaker ethnicity on (1) accentedness, (2) intelligibility and (3) comprehensibility. The latter dimension has not been investigated within the context of the RLS hypothesis and the exemplar theory framework, so this dissertation project was designed to test whether either of the two theories' competing predictions is borne out in the data.

introspection, but do not agree that the associations assessed by implicit measures are necessarily unconscious" (p. 5). This notion was reiterated by Pharao and Kristiansen (2019).

[9] Some language attitude researchers may still consider Likert scales to be an indirect measure, but in the implicit memory tradition within the field of social psychology (and for the purposes of this dissertation) Likert scales are seen as a direct measure because they allow introspection. See Pantos (2019) for a more in-depth explanation of the different uses of this terminology.

Chapter 3.

Experiment 1 with photographs in Vancouver

Various studies have provided evidence that it takes native speakers longer to process foreign-accented speech than native-sounding speech, whether in a sentence verification task (Adank, Evans, Stuart-Smith, & Scott, 2009; Munro & Derwing, 1995), a verbal repetition task (Perry, Mech, MacDonald, & Seidenberg, 2018; Weil, 2003), a word identification task (Floccia, Butler, Goslin, & Ellis, 2009; Floccia, Goslin, Girard, & Konopczynski, 2006), or a pronunciation error task (Schmid & Yeni-Komshian, 1999). This increased processing effort for non-native speech should not come as a surprise. After all, non-native speech has far more variation in phonological patterns, higher error rates, and greater degrees of overall unpredictability compared to native speech (Adank et al., 2009; Clarke & Garrett, 2004; Romero-Rivas, Martin, & Costa, 2015).

Listeners do indeed seem to process native and non-native speech differently (Grey, Schubel, McQueen, & van Hell, 2019; Grey & van Hell, 2017; Lev-Ari et al., 2018; Romero-Rivas et al., 2015). For instance, listeners who believe that they are listening to a non-native speaker adjust their expectations about the likelihood of errors to optimize processing (Romero-Rivas et al., 2015). ERP research has also demonstrated that listeners only showed surprise to a syntactic error that had been made by a native speaker, whereas this neural response was absent when the error was made by a non-native speaker (Caffarra & Martin, 2019; Hanulíková, van Alphen, van Goch, & Weber, 2012).

To increase our understanding of listener attitudes in social interactions, it is important to study the influence of sociolinguistic expectations on language processing. One way to do this is by measuring how long it takes listeners to respond to stimuli. The underlying assumption of these kinds of response time studies is that longer response times (RTs) reflect an additional processing load (Maas & Mailend, 2012; Pantos, 2019). While we know that listeners try to infer as much information about a speaker as they can to facilitate effective processing, the degree to which *visual information* may have an influence on reaction times has barely been investigated. The aim of this experiment was therefore to examine whether seeing an Asian or White face differently affects the time required by listeners to decide on the truthfulness of a statement spoken in a native or

foreign accent. The methodology is largely a replication of the sentence verification task that was used by Munro and Derwing (1995) in a seminal study that is now frequently cited as evidence for increased processing difficulty of foreign-accented speech.

Following Babel and Russell (2015) this study was conducted in the multi-ethnic and linguistically diverse city of Metro Vancouver, where 48.9% of the population identifies as a visible minority, and 42.9% of the population speaks a language other than English as their mother tongue (Statistics Canada, 2017c). Since many Metro Vancouver's immigrants (78.9%) were born in Asia, it is very likely that most Vancouverites will conjure up the image of an Asian face when asked to think of a 'typical immigrant' in Vancouver. There are however also many second- and third-generation Asian Canadians who were born and raised in Canada, and thus speak English as their native tongue. Seeing that Vancouverites interact daily with both Asian (non-native speaking) immigrants and Asian Canadians, there is a possibility that they have learned to base their linguistic expectations less on a speaker's ethnicity, since this type of indexical information does not provide reliable information about the speaker's nativeness within the context of Metro Vancouver. This experiment's social context stands in sharp contrast with where Rubin and his colleagues conducted their experiments, as Rubin (1992) himself acknowledged that "the geographic region in which this study was conducted [i.e., the area around the University of Georgia] has a relatively low proportion of nonnative English speakers and of Asians. These students are exposed to relatively little nonnative accented speech in their daily affairs" (p. 529).

As was mentioned by Babel and Russell (2015), even though there are many ethnically Asian native speakers of English living in Vancouver, Asian Canadians are often still erroneously expected to be non-native speakers. While Babel and Russell uncovered that listener expectations–as informed by visual speaker ethnicity–appeared to affect intelligibility, this experiment investigates whether these listener expectations affect comprehensibility as well. This will be tested on a speech processing level by measuring reaction times to various manipulated guises.

3.1. Methods Experiment 1A

3.1.1. Stimulus sentences

A speeded sentence verification task ('Experiment 1A') was created to determine the effect of accent and face on response latencies. A list of 60 statements (30 true and 30 false) was created, which was largely adapted from Munro and Derwing's (1995) original stimulus list of forty sentences that had already been vetted for potential ambiguities. The twenty new sentences added to this list were informally evaluated by two individuals on potential ambiguities. Sentences were kept as short as possible to avoid working memory limitations interfering with processing speeds (as observed in Floccia et al., 2006). The resulting sixty sentences were all easily identifiable as either true or false for the average North American listener (e.g., *a monkey is a kind of bird; a face can show expression*). After the stimulus sentences had been recorded by four speakers, four sentences were removed from the dataset due to intelligibility issues identified by the experimenter. The remaining fifty-six sentences (28 true, 28 false) were all single-clause sentences that ranged between four and nine words, with a mean length of 6.0 words. The sentences are listed in Appendix A. Since various studies have uncovered differences in identification performance between high and low-frequency words (Bradlow & Pisoni, 1999; Mullenix, Pisoni & Martin, 1989; Nygaard & Pisoni, 1998), this study's sentences were carefully constructed to only contain the most commonly used words in English based on frequency count. These words were taken from a list of 1000 words with the highest frequency count in English based on written corpora, which included both content and function words (Fry, 2011).

3.1.2. Speakers

Two speakers were recruited for each of the two accent categories to help address the concern that any observed RT differences could otherwise be talker-independent (Bradlow & Bent, 2008). While more than two speakers per accent would have made the results even more generalizable, practical limitations made this impossible for the scope of this experiment. Two female speakers of Canadian English (aged 41 and 67) provided the two native-accented voices. They were coded as 'Native1' and 'Native2,' respectively. Since listeners with more exposure to accents have been found to perform better than

those with less experience (e.g., Floccia et al., 2009; McGowan, 2015), a conscious effort was made to pick a foreign accent that would be relatively unfamiliar to the listeners. To this end, Japanese-accented speech was chosen, since Japanese is only spoken as a mother tongue by 0.7% of the total population in the metropolitan area in which the study was conducted (Statistics Canada, 2017a).

The speaker who provided the voice dubbed 'Foreign1' was a 28-year-old woman who had moved from Japan to Canada six months prior to recording. 'Foreign2' was an amalgamation of the most intelligible recordings of four twenty-year-old Japanese women who had all been studying at a Japanese university at the time. This route had to be taken because none of the four speakers who resided in Japan had produced a stimulus set that was fully intelligible.[10] The Canadian speakers and one of the foreign-accented speakers ('Foreign1') were recorded at a Canadian university, while 'Foreign2' was recorded at a Japanese university. The two Japanese-accented 'voices' received different accentedness scores and comprehensibility ratings (see section 3.4.3), but Analysis IV demonstrates that they did not differ in terms of actual comprehensibility, as they did not elicit different response times.

To select the most intelligible tokens for each sentence, two native English listeners independently assigned perceived intelligibility ratings to the four Japanese speakers' utterances on a 9-point scale. Based on these scores, the best rated rendition was chosen for each sentence. Not all speakers were equally represented (NJ1 = 12; NJ2 = 3; NJ3 = 21; NJ4 = 20). In addition to the test stimuli, a 62-year-old native female Canadian English speaker provided the recordings of ten practice statements that were played to all participants before the test block.

3.1.3. Task Design

Each sentence was pseudo-randomly paired with either a single image of a young East Asian woman, a young Caucasian woman, or with a black fixation cross on a white background. The two photographs of the women were taken from the Chicago Face Database, which is a freely accessible stimulus set of faces (Ma, Correll, & Wittenbrink,

[10] The Canadian voice donors sounded noticeably older than the Japanese voice donors, which may have added an additional, unintentional incongruent layer to the analysis.

2015). Appendix B shows the three images that were used. To ensure a balanced design in which each participant heard all 56 sentences only once and where the Asian and White face were shown equally often, thirty-two unique stimulus sets were created. Within each stimulus set, the presentation of the stimuli was randomized. Each stimulus set contained 28 true and 28 false statements, which were equally divided among the four speakers. Each speaker was represented 14 times in each stimulus set, of which six were played alongside the Asian face, another six with the White face, and the remaining two with a fixation cross. This design is presented visually in Figure 3.1.

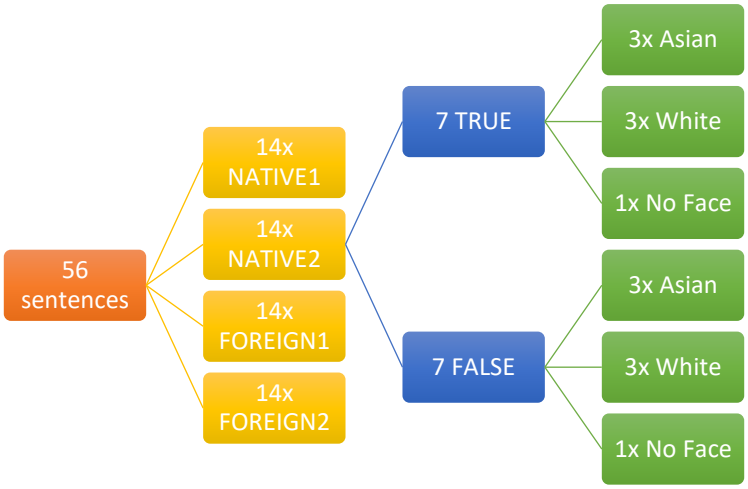


Figure 3.1 Breakdown of the *Face–Voice* combinations for each participant

The speakers were all recorded individually in a sound-attenuated booth with high-fidelity audio equipment. They were given the list of sentences and were asked to read it aloud in its entirety. They each read the stimulus list twice consecutively, with a break in between. When there were noticeable hesitations or pronunciation errors, they were asked to read these items a third time. The experimenter helped with the pronunciation of a few words for one of the foreign-accented speakers ('Foreign1'), but only when it was apparent that the mispronunciation was largely due to orthography.

Each recording was trimmed at a zero crossing in Praat (Boersma & Weenink, 2018) to ensure that there was no leading or trailing silence. Utterance-final fricatives and vowels were cut off at the earliest possible moment without this negatively impacting intelligibility as judged by the experimenter. The total durations of the speakers' utterances were not modified, which meant that each sentence had varying durations depending on the talker. As expected, the non-native utterances took longer overall (Mforeign1 = 2483 ms;

Mforeign2 = 2604 ms) than the native utterances (Mnative1 = 1886 ms; Mnative2 = 2022 ms). All audio files were normalized using Audacity (Audacity Development Team, 2018) and were presented to the listeners with no background noise.

3.1.4. Listeners

Thirty-two listeners were recruited by means of convenience sampling. The listeners were all self-reported native speakers[11] of English (21 female, 11 male) and were all Caucasian. Their respective native English varieties were Canadian (20), British (7), American (2), South African (2), and Irish (1). Two listeners had an Asian background, while the other thirty were Caucasian. Following previous research[12], no restrictions were placed on the ethnic backgrounds of the participants. A recent study did observe that a listener's ethnic group affiliation (EGA) could exert an influence on how they rate other speakers' degree of accentedness (Tekin, 2019), but in order to keep the number of variables manageable in this experiment–and the other two described in Chapters 3 and 4–the decision was made to not take EGA into account.

When asked how frequently they interacted with Japanese-accented speakers, eleven participants said almost never, thirteen participants said monthly, six indicated they interacted with Japanese speakers on a weekly basis, and two said daily. They all reported normal hearing and were thanked for their participation but received no financial compensation. Participants were asked to select the age range that applied to them, with the youngest category being the 21-25 age range, and the oldest category being the 71-75 age range. See Figure 3.2 below.

[11] For the purpose of this study, I will use Hanulíková's (Hanulíková, 2018) definion of native and non-native speakers: "I use the terms native and nonnative speaker to distinguish between a person who grew up speaking a given local variety and uses it regularly and a person who acquired a foreign language at school age or later" (p. 8).

[12] Zheng and Samuel (2017) were the only ones to exclude "East Asian participants to avoid a potential effect of own-race preferences when presented with stimuli that contained an East Asian face" (p. 1843).

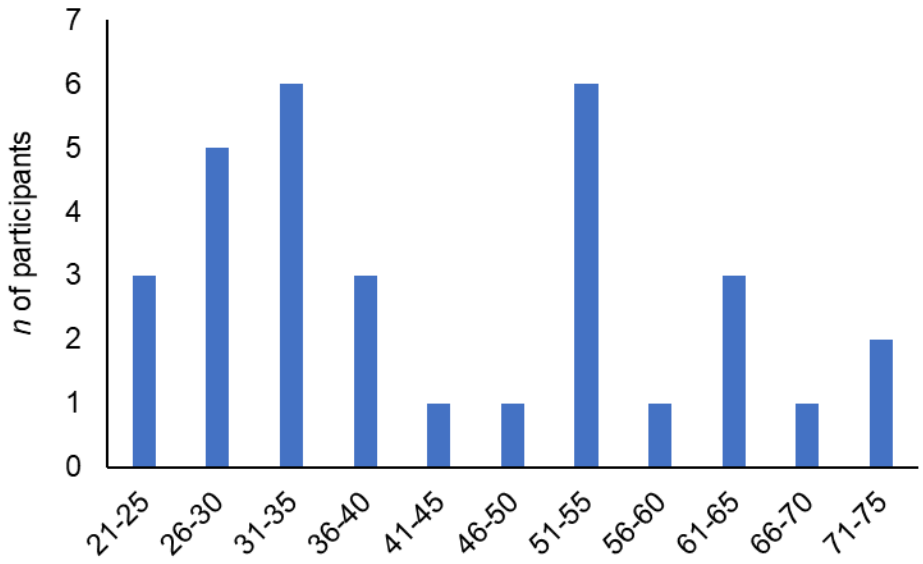


Figure 3.2 Age distribution of participants

3.1.5. Procedure

Participants were tested individually in a silent room. After the experiment procedure had been explained to them, they were asked to sign an informed consent form and were seated behind an iPad (running on iOS 11.2) that had been propped up on a stand in landscape mode. A pair of noise-cancelling headphones (Model: Sony WH1000XM2/B XB) was connected to the iPad, and the participants were told that they were free to adjust the volume if needed during the practice session. The audio files and images were presented with the *Paradigm for Mobile* app, which is a millisecond accurate stimulus presentation system for iPhone and iPad (Perception Research Systems, 2007). The participants were informed that this was a two-alternative forced choice task where they had to identify each sentence as true or false as quickly as possible without sacrificing correctness. They were furthermore instructed to use their dominant index finger when selecting responses, and to not move the iPad from its stand. Finally, the researcher (who was present throughout the entire experiment) emphasized to the participants that they had to pay attention to the faces during the experiment.

The participants first completed a short ten-item practice session on sentences that had been read aloud by a 62-year-old female Canadian English speaker. The sentences were presented together with a fixation cross. The image and audio were always presented simultaneously, along with two large touch response boxes on the

bottom of the screen. The left button was colored green and featured the word TRUE while the right button was colored red and featured the word FALSE. The order of the buttons did not change throughout the experiment. After the audio had ended, the image remained on the screen until a button was selected or the trial timed out. If the system did not detect a response within seven seconds after a stimulus had been presented, the next stimulus was shown. After completing the practice session, a short announcement was shown informing the participants that when they were ready, they could touch the 'proceed' button to start the experiment. The entire procedure took an average of six minutes.

After judging all fifty-six sentences, each participant answered some background questions about their age, their English variety, and their exposure to Japanese-accented speech. The latter question was included because evidence suggests that familiarity with an accent sometimes facilitates understanding (Bradlow & Bent, 2003; Gass & Varonis, 1984; Thompson, 1991; Weil, 2001), and lowers the perceived degree of accentedness. Because *Paradigm* did not allow text responses, listeners were asked to choose the relevant age range from a series of touch response boxes. To convert participant age from a categorical to a numerical variable, the middle value was selected from each five-year age range to represent the participants' approximate age.

Paradigm recorded each listener's response times to the millisecond starting from the moment the stimulus started playing. Since the four speakers' renditions of the same sentence varied in duration, the audio file's duration was therefore subtracted from the total reaction time to calculate a post-sentence reaction time. Besides response times, *Paradigm* also recorded each listener's response (true/false) to each sentence and compared this to the correct response. A no-response and wrong response both received a value of 0, whereas a correct response received a value of 1. The two South African listeners' responses to the sentence #33 'August is a winter month' were removed, as this sentence is true for people living in the Southern hemisphere, but false for Canadians.

3.2. Methods Experiment 1B

Even though the sentences had been presented without any added background noise, many participants commented that some of the sentences were completely unintelligible to them. Therefore, a follow-up study ('Experiment 1B') was conducted to

measure the perceived degrees of accentedness and comprehensibility for the accented voices Foreign1 and Foreign2, respectively. The main aim was to cross-reference these collected ratings with response times, to see if perceived degrees of comprehensibility and accentedness could predict response times.

3.2.1. Stimuli

All the Japanese-accented true/false sentences used in Experiment 1A were included in Experiment 1B (*n* = 112). An additional twelve Canadian-accented utterances were added to serve as baseline ratings. To measure intra-rater reliability, twenty-six utterances were repeated once. This resulted in a total of 150 tokens.

3.2.2. Listeners

The ten listeners who participated in this Audio-Only rating task were drawn from the same participant pool who completed the Audio-Visual experiment. The time between completing the Audio-Visual experiment and the Audio-Only rating task was at least a month.

3.2.3. Procedure

A non-speeded rating task was created in *Praat* (Boersma & Weenink, 2018). Listeners were given noise-canceling headphones (Model: Sony WH1000XM2/B XB) and were seated behind an 11” Macbook Air laptop with a separate mouse and mousepad. They were allowed three replays of each item, but were encouraged to respond quickly and intuitively, rather than to overthink their responses. There was an enforced break halfway through.

The listeners were instructed to assign perceived accentedness and comprehensibility ratings on nine-point Likert scales to all the utterances they heard (1 = no accent, 9 = very strong accent; 1 = very easy to understand, 9 = very difficult to understand). Both scales were presented simultaneously for each item–accentedness on the top, and comprehensibility on the bottom. The whole task took between 20 and 35 minutes. Listeners were financially remunerated for participating.

3.3. Predictions

Previous research has focused almost exclusively on how visual speaker information affects intelligibility and accentedness, while ignoring the third dimension of comprehensibility.[13] This experiment addresses this research gap by investigating the effect of visual speaker ethnicity on speech processing through response latencies. If the 'ethnicity effect' observed by Babel and Russell (2015) and McGowan (2015) on accentedness ratings and intelligibility scores can be extended to comprehensibility too, we would expect to see worse performance (i.e., longer response times) in the incongruent face–voice conditions compared to the congruent ones. Yet if the listeners have more difficulty processing the speech of the Asian guises compared to the Caucasian ones–regardless of their accent–then the data would be more in line with the reverse linguistic stereotyping hypothesis.

Within the context of this study (listeners in Vancouver), the most incongruent pairing is expected to be a White guise speaking with a Japanese accent. The other unexpected pairing in this experiment is assumed to be an Asian guise speaking with a native English accent. These predictions may come as a surprise to some readers, especially considering that there are so many Asian Canadians living in the Greater Vancouver area who speak English as their mother tongue. Yet this study's assumptions about who Vancouverites will most likely expect to be native speakers are not chosen arbitrarily; instead, they are informed by previous observations made in two related studies conducted in Vancouver (Babel & Mellesmoen, 2019; Babel & Russell, 2015), in addition to Canadian Census data on immigrant groups in Vancouver, and the general findings discussed in section 2.2.3 that 'the prototypical native speaker' is generally still expected to be Caucasian in appearance.

Knowing that non-native speech has far more variation in phonological patterns, higher error rates, and greater degrees of overall unpredictability compared to native speech (Adank et al., 2009; Clarke & Garrett, 2004; Romero-Rivas, Martin, & Costa, 2015), and that "research on processing has demonstrated that increased variability in the signal places a higher demand on the normalization process at the time of perception"

[13] Hanulíková (2018) appears to be the only one to have investigated the effect of speaker ethnicity on a comprehensibility Likert scale.

(McLennan & Luce, 2005, p. 306), we would expect foreign-accented speech to take longer to process than native-accented speech, just like Munro and Derwing (1995) had found.

The older listeners are furthermore expected to have overall longer response times than the younger listeners because of slower cognitive processing. Finally, since there were noticeable differences in accentedness between the native Canadian and foreign-accented voices, the Japanese-accented voices are predicted to (1) receive higher accentedness ratings, (2) take longer to process, and (3) have lower accuracy scores compared to the Canadian-accented voices (as previously reported by Munro and Derwing (1995) and Adank et al. (2009).

3.4. Results

3.4.1. Analysis I: Response times

The listeners' reaction times to the true and false statements in each of the face–voice pairings were analyzed using the open source statistical package *R* (R Core Team, 2014). A linear mixed effects model was used to examine the relationship between the fixed effect variables *Voice* (native, accented), *Face* (Asian, White, no face), *Veracity* (true statement, false statement), and the participants' approximate *Age* on response times. *Age* was included as a factor, as there is evidence suggesting that (1) older adults tend to be more prejudiced than young adults (Gonsalkorale, Sherman, & Klauer, 2009; Stewart, von Hippel, & Radvansky, 2009; Von Hippel, Silver, & Lynch, 2000) and that (2) older listeners have more difficulty perceiving speech than younger listeners, possibly due to differences in hearing acuity and working memory capacity (Ingvalson, Lansford, Fedorova, & Fernandez, 2017). Due to time constraints, listener experience with Japanese-accented speech was not further analyzed after a cursory analysis did not suggest any significant influence. Since observations within a participant are likely related–which would violate the assumption of independence–the term *Participant* was included to control for random participant effects. This variable accounts for the dependencies that likely exist within the observations for a given participant. Treating participants as random effects allows the results to be generalized to the entire population, and not just to the thirty-two listeners who participated in this study. It is important to note

that the model was only fit to the response times of correctly identified sentences; RTs to incorrectly identified sentences (*n* = 35), timed-out sessions (*n* = 83) and instances where participants had prematurely pushed a button before the sentence had ended *(n* = 39) were not included in the analysis.

A fitted vs. residuals plot and a Q-Q plot showed that the residuals in the initial model violated both the constant variance and normality assumptions of regression, so a natural log transformation was applied. This transformation addressed the variance violation, but the normality assumption still appeared to be violated. Yet because ANOVAs are fairly robust to violations of the normality assumption, it was decided to continue with the analysis. There was a total of 19 observations with an absolute value for residuals of < 2.5. Generally, these outliers had much lower response times than the rest of the data. Because a re-run of the analysis without the outliers did not render different results, it was decided to keep the outliers in the analysis. Figure 3.3 below shows the spread of RTs without the 19 identified outliers. Only the RTs to correct responses were plotted.

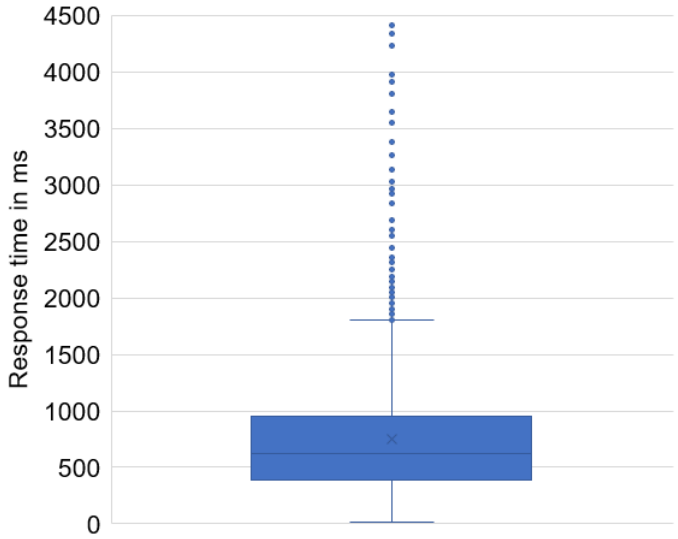


Figure 3.3 Boxplot showing the spread of raw response times in ms

Several other models were examined, but since these none of these models had significant interaction terms, all interaction effects were removed from the final model except for the interaction between *Voice* and *Face*, as it was of interest to this analysis. The lmerTest package (Bates, Mächler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2017) used dummy variable recoding to produce beta estimates. For this particular analysis, the following coding scheme was used:

Table 3.1 Variable Coding Scheme

$X_1 = 1$	if sentence *Veracity* is true	$X_1 = 0$	if sentence *Veracity* is false
$X_2 = 1$	if *Face* is No Face	$X_2 = 0$	if *Face* is White or Asian
$X_3 = 1$	if *Face* is White	$X_3 = 0$	if *Face* is Asian or No Face
$X_4 = 1$	if *Voice* is native	$X_4 = 0$	if *Voice* is accented
$X_5 =$	Age of participants * estimated regression coefficient for Age		

Table 3.2 Estimates of fixed effect regression coefficients for the linear mixed effects model with corresponding *p*-values

	Estimate	SE	df	t	Pr(>|t|)	Sig
(Intercept)	6.287	0.202	33	31.109	0.000	***
Veracity_True	-0.1476	0.038	1595	-3.9	0.000	***
Face_NoFace	-0.0547	0.083	1596	-0.655	0.513	
Face_White	-0.0016	0.059	1596	-0.027	0.979	
Voice_Native	-0.2942	0.058	1595	-5.08	0.000	***
Age	0.0066	0.004	30	1.538	0.135	
FaceNoFace : VoiceNative	-0.0596	0.116	1595	-0.514	0.607	
Face_White : Voice_Native	-0.0373	0.082	1595	-0.456	0.649	

Note: *** p <0.001

Random effects

Groups	Variance	Std.Dev.
Participant (intercept)	0.1261	0.3551
Residual	0.5839	0.7641

Table 3.2 above shows a summary of the model output. After fitting the above model in *R*, an ANOVA was carried out on the fitted model using Kenward-Roger's method with Type 3 Sums of Squares. The results are in Table 3.3 below. As can be seen, both *Veracity* and *Voice* were highly significant. There was, however, no evidence found of an interaction effect between *Face* and *Voice*, or main effects for *Face* or listener *Age*.

Table 3.3 Type III Analysis of Variance Table with Kenward-Roger's method

	Sum of Squares	Mean Squares	NumDF	DenDF	*F*	*p*
Veracity	8.88	8.88	1	1595	15.21	< .001
Voice	33.36	33.36	1	1595	57.13	< .001
Face	1.24	0.62	2	1596	1.06	0.35
Age	1.38	1.38	1	30	2.37	0.13
Voice * Face	0.21	0.10	2	1595	0.18	0.84

The interaction plot below provides additional visual support for the observed significant main effects of *Veracity* and *Voice*, with the Japanese-accented voices seemingly taking longer to respond than the Canadian-accented voices, and the false sentences in turn taking longer than the true ones. The plot also illustrates that response times were unaffected by *Face–Voice* (in)congruence, as there were virtually no differences between response times for the congruent or incongruent *Face–Voice* pairings[14]. Response times also appear to be slightly faster in the Audio-Only condition for both accents and sentence types (true/false). Note that the graph below depicts the back-transformed response times and their standard errors. A graph visualizing the raw data (i.e., mean RTs) has been included in Appendix C.

[14] Congruent = Japanese-accented + Asian / Canadian English + White;
Incongruent = Japanese-accented + White / Canadian English + Asian

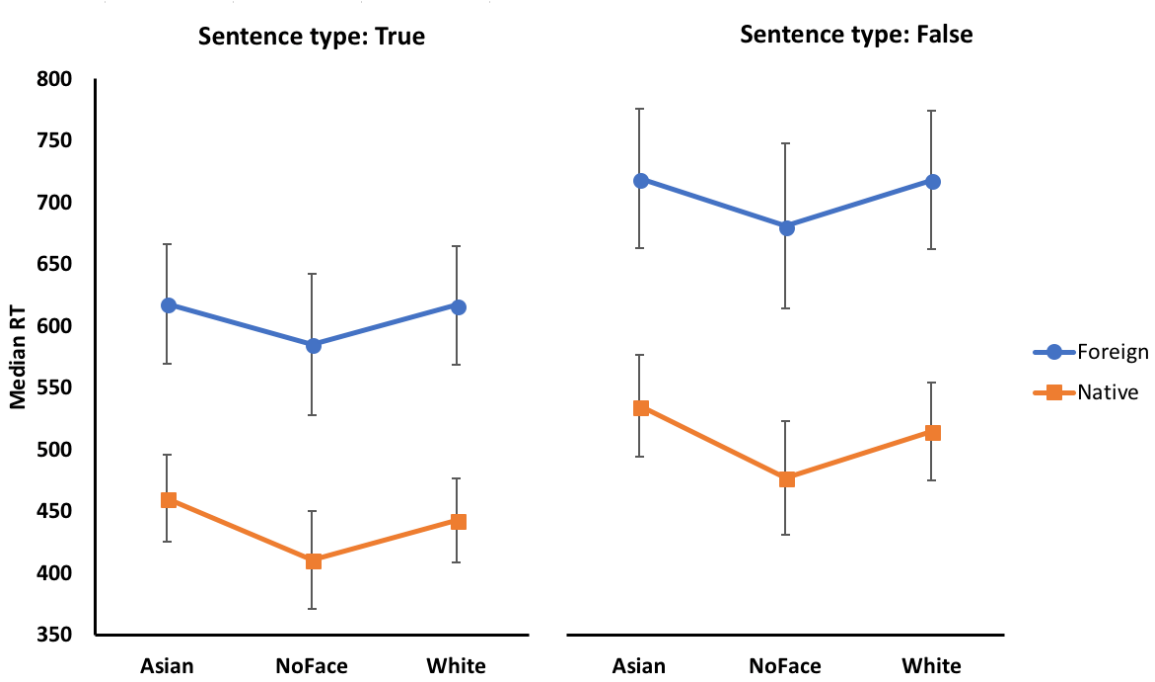


Figure 3.4 Median RTs with SE bars for all levels of *Face* and *Veracity*

Since the ANOVA had indicated that there were significant main effects of *Voice* and *Veracity*, the emmeans package in *R* (Lenth, 2019) was used to generate the estimated mean RTs of the two levels for each variable using the results and coefficients of the Lmer model fit. Note that the resulting log(RT) and confidence intervals were back-transformed into median RTs for easier interpretation. Results show that, on average, Japanese-accented voices tended to take longer to process to than native voices (difference between the medians: 181.9 ms), and that false statements took longer to process than true statements (difference between the medians: 82.0 ms).

Table 3.4 Back-transformed estimated median response times in ms for *Voice*

				95% Confidence Interval	
Voice	Median RT	SE	df	Lower	Upper
Native	471	32.77	38	409.25	542.39
Foreign	653	45.76	39	566.75	752.46

Table 3.5 Back-transformed estimated median response times in ms for both levels of *Veracity*

				95% Confidence Interval	
Veracity	Median RT	SE	df	Lower	Upper
True	515	35.58	37	447.94	592.60
False	597	41.21	37	519.23	686.81

Then, since the ANOVA had yielded significant effects for both *Voice* and *Veracity*, a follow-up Tukey-Kramer test was carried out using the Kenward-Roger degrees-of-freedom approximation to test whether any of the differences between the four individual *Voice–Veracity* pairings would achieve significance, as this was of interest to this study. The results shown in Table 3.6 below indicate that there were in fact significant differences between the four pairings, but that there was no interaction between them, as shown in the ANOVA; listeners responded the fastest to true sentences spoken in a Canadian accent, followed by false native-spoken statements, and true foreign-accented statements. Finally, false sentences spoken with a Japanese accent had the longest response times. In other words, responses to the native voices were faster than those to non-native voices, and within each accent, the false sentences entailed longer RTs than true sentences.

Table 3.6 Back-transformed estimated median response times in ms for each *Voice–Veracity* pairing

					95% Confidence Interval		
Voice	Veracity	Median RT	SE	df	Lower	Upper	Group*
Native	True	438	31.54	43.76	378.44	506.05	1
Native	False	507	36.57	43.84	438.62	586.58	2
Foreign	True	607	44.06	45.14	524.02	702.13	3
Foreign	False	703	50.99	44.84	607.51	813.64	4

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Discussion

Analysis I showed no indication that guise ethnicity affected response times, which could be interpreted as evidence that neither (in)congruence, nor listener bias affected comprehensibility. It is however important to acknowledge that this lack of an effect may have been the result of listeners ignoring the guises' purported ethnic identities entirely, as they were not consistently paired with a single voice and were therefore likely perceived

as irrelevant. Another methodological decision that may have influenced the results is the fact that the sentences were presented without added background noise, which may have resulted in the task being too easy.

Findings furthermore showed that true sentences were responded to faster than false statements, but contrary to expectation, there were no differences in RTs between the younger and older participants. The accent in which sentences were spoken did appear to make a difference, with listeners having overall faster response times to Canadian-accented utterances than to Japanese-accented ones, which strongly suggests that the Canadian speakers were easier to understand than the Japanese-accented speakers.

3.4.2. Analysis II: Probability of correct sentence identification

This next analysis focused on the number of correctly identified sentences. A mixed logistic regression was fit to the data to investigate the effect of the presented face–voice combination on the probability of listeners correctly judging a statement. All observations where the response time was negative or where the participant did not give a response were filtered out of the analysis (n = 76, which amounts to 4.2% of the total number of responses). In contrast to the previous analysis, this model did not contain a random error term. The other variables are the same as the previous analysis, with the omission of the *Age* variable due to convergence issues. Since there was no significant effect of *Age* (as tested in a second model), the removal of this fixed effect was justifiable.

An ANOVA using Type 3 sum of squares was carried out on the fitted model to determine if any variables significantly affected the likelihood of listeners correctly judging the sentences' truth value. Table 3.7 below shows the results. Once again, as can be seen from this table, the only statistically significant effects are the main effects for *Voice* and *Veracity*.

Table 3.7 Analysis of variance using type III sum of squares

	Chi-Squared Statistic	Degrees of Freedom	*p*
Face	2.28	2	0.321
Voice	26.16	1	< .001
Veracity	5.73	1	0.017
Face * Voice	3.29	2	0.193

Additional evidence against an interaction effect is provided by the interaction plot below (Figure 3.5). Despite appearances, no significant interaction between *Veracity* and *Voice* was found.

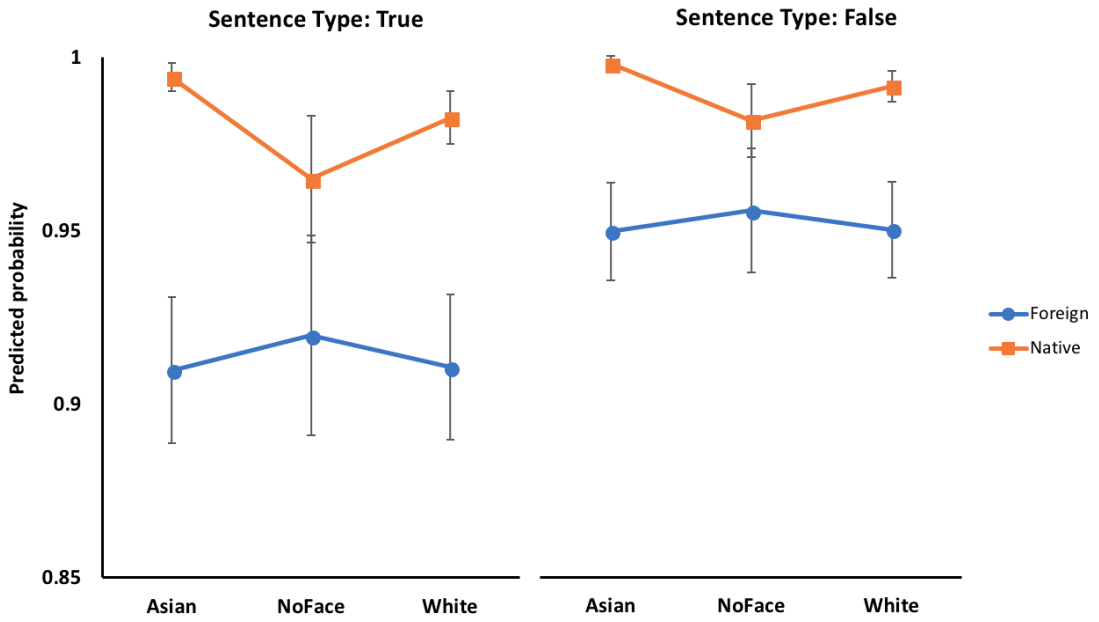


Figure 3.5 The predicted probabilities of listeners correctly judging a statement on all *Voice–Veracity* levels

Post-hoc Wald tests were used to further examine the main effects of *Veracity* and *Voice* on the estimated probabilities of participants correctly verifying a sentence as true or false. Table 3.8 and Table 3.9 below show the estimated probabilities averaged over *Voice* and *Veracity* individually. They provide statistically significant evidence that (1) participants were more likely to correctly judge the veracity of a sentence that had been spoken in a Canadian accent compared to a Japanese accent, and (2) that false statements were more likely to be correctly identified than true statements. This effect of *Veracity* disappeared when the group means of the four individual *Voice–Veracity* combinations were compared (see Table 3.10), but this is most likely due to the more stringent *p*-values that are required in multiple comparisons.

Table 3.8 The effect of *Voice* on correct judgment probability

			Asymptotic Confidence Interval		
Voice	Probability	SE	Lower	Upper	Group*
Foreign	0.935	0.013	0.905	0.956	1
Native	0.987	0.004	0.976	0.993	2

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Table 3.9 The effect of *Veracity* on correct judgment probability

			Asymptotic Confidence Interval		
Veracity	Probability	SE	Lower	Upper	Group*
True	0.962	0.009	0.940	0.976	1
False	0.978	0.006	0.963	0.987	2

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Table 3.10 Correct judgment probability for each *Voice–Veracity* combination

				Asymptotic Confidence Interval		
Voice	Veracity	Probability	SE	Lower	Upper	Group*
Native	True	0.983	0.006	0.967	0.991	1
Native	False	0.990	0.003	0.981	0.995	1
Foreign	True	0.915	0.018	0.874	0.944	2
Foreign	False	0.951	0.012	0.920	0.970	2

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Discussion

Analysis II investigated whether the likelihood of listeners correctly identifying a sentence as true or false was dependent on *Veracity*, *Face*, *Voice*, or participant *Age*. It was once again found that neither guise ethnicity *n*or listener age influenced this likelihood. What did affect the probability of a correct identification was the accent of the speakers, as listeners were more likely to correctly judge the truthfulness of sentences that had been spoken by the Canadian speakers. Listeners were also more likely to correctly identify false statements compared to true ones. Since Analysis I had found that listeners required more time to respond to false statements, these findings taken together suggest a correlation between the time participants take to respond to a statement and the likelihood of correctly identifying a sentence's veracity.

3.4.3. Analysis III: Predicted RTs based on comprehensibility and accentedness

This next analysis investigated if the accentedness and comprehensibility ratings that had been collected in the Audio-Only rating task (Experiment 1B) could serve as predictors for response times. To calculate inter-rater reliability, the repeated items' scores were averaged. Cronbach's alpha was .93, indicating very high correlation among the judges. Intra-rater reliability was measured by calculating the differences between each listener's first and second rating of the twenty-six repeated sentences. The results presented in Table 3.11 and Table 3.12 below show that all users consistently assigned identical ratings to repeated Canadian-accented sentences, namely 'not accented,' and 'very easy to understand.' Table 3.13 shows the mean scores assigned to the two foreign-accented and two native-accented voices on a nine-point scale.

There were small rating inconsistencies within users for the two foreign voice guises, but the standard deviations of these differences were between 1 and 2, which suggests that people were generally on target. Additionally, the confidence intervals for the means of both metrics show that the mean differences were not significantly different from zero, which provides confirmatory evidence that, on average, scores were consistent as well. It furthermore demonstrates that there is no indication of some sort of correction factor occurring where users tended to give the same sentence a higher or lower rating the second time around.

Table 3.11 Differences in mean accentedness ratings for each voice

			95% Confidence Interval	
Voice	Mean_Acc	sd	Lower	Upper
Native1	0	0	0	0
Native2	0	0	0	0
Foreign1	-0.02	1.45	-0.29	0.26
Foreign2	0.14	1.23	-0.10	0.37

Table 3.12 Differences in mean comprehensibility ratings for each voice

			95% Confidence Interval	
Voice	Mean_Comp	sd	Lower	Upper
Native1	0	0	0	0
Native2	0	0	0	0
Foreign1	0.19	1.44	-0.08	0.47
Foreign2	-0.03	1.72	-0.35	0.30

Table 3.13 Actual mean ratings and standard deviations of the two foreign-accent voices (9-point scale)

Voice	Mean_Acc	sd	Mean_Comp	sd
Native1	1	0	1	0
Native2	1.04	0.19	1	0
Foreign1	7.21	1.36	5.11	1.60
Foreign2	5.99	1.58	3.6	1.35

The ten listeners' ratings were averaged to calculate mean accentedness and comprehensibility scores for each *Veracity–Voice* combination. Remember that all the foreign-accented sentences had received individual ratings, while only 16 of all the native-spoken sentences had been rated to serve as a baseline. Since most of the native *Veracity–Voice* combinations had been assigned the lowest score of 1 in both categories (i.e., no accent; easy to understand), this score was extended to all the remaining native *Veracity–Voice* pairings as well. This meant that all native-spoken sentences were assigned mean comprehensibility and accentedness scores of 1, while all foreign voices received a mean score of >1. As such, the resulting continuous variables *accentedness* and *comprehensibility* are essentially finer-grained versions of the categorical *Voice* variable in that they capture more detailed information about the individual speakers' levels of accentedness and comprehensibility.

Before the two new variables could be fit into a new linear mixed effects model, the correlation coefficient had to be calculated between these two new variables. Not only is it redundant to include two highly correlated independent variables in the model, but doing so would also cause the *p*-values calculated for these independent variables' regression coefficients to be unstable from sample to sample. The correlation between *accentedness* and *comprehensibility* was 0.945, which indicates that sentences with higher foreign-accentedness scores were also rated as more difficult to understand.

Because of this very high correlation coefficient, only one of the two variables' scores could be used in the regression model. It was decided to replace the categorical *Voice* variable with the continuous variable *mean_comp*, which represented a sentence's average comprehensibility rating.

As in Analysis I, other models were explored, including interaction terms for *Age* and *Veracity*, but none of these interactions were significant. The same outliers (n = 19) that affected the normality assumptions in Analysis I also appeared in this analysis, in addition to one new outlier. Since their removal did not change the overall outcome, the outliers were kept in the analysis. An ANOVA was carried out on this model using Kenward-Roger's method with Type 3 Sums of Squares. As can be seen below in Table 3.14, like the first analysis, only *Veracity* and *mean_comp* were significant.

Table 3.14 Type III Analysis of Variance Table with Kenward-Roger's method

	Sum Sq	Mean Sq	NumDF	DenDF	*F*	*p*
Veracity	5.271	5.271	1	1595	9.417	0.002
Face	0.484	0.242	2	1596	0.432	0.649
Mean_comp	60.436	60.436	1	1596	107.974	< .001
Age	1.312	1.312	1	30	2.344	0.136
Face * Mean_comp	0.088	0.044	2	1595	0.078	0.925

The back-transformed estimate of the regression coefficient for comprehension scores was 0.121, with R^2 = 0.1497, meaning that a one unit increase in mean comprehension scores is associated with a multiplication of log(response times) by 0.121. Put differently, after all other variables are controlled for, an individual's median response time for a sentence with mean_comp = 2 is estimated to elicit a response time that is 0.121 times longer than the median response time for a sentence with mean_comp = 1. The fact that the regression coefficient is a positive number indicates that statements with larger mean comprehensibility scores (i.e., rated as 'harder to understand') are also associated with longer log(response times). A possible interpretation is that impressionistic Likert scale ratings and RTs both share an underlying influence. If so, insight into processing difficulty can be gained through both explicit scalar ratings and covertly by means of RTs. However, causation cannot be established from the data collected in this study.

Figure 3.6 below shows a scatter plot of the mean comprehensibility scores in relation to log(response times). For this figure, all observations where an individual was

presented with a native voice were removed, as all these observations had a mean comprehensibility score of 1 (i.e., 'easy to understand'). Only correct responses were included. Once again, higher *mean_comp* scores meant that the sentences were rated as more difficult to understand. Note that log(response times) were modelled.

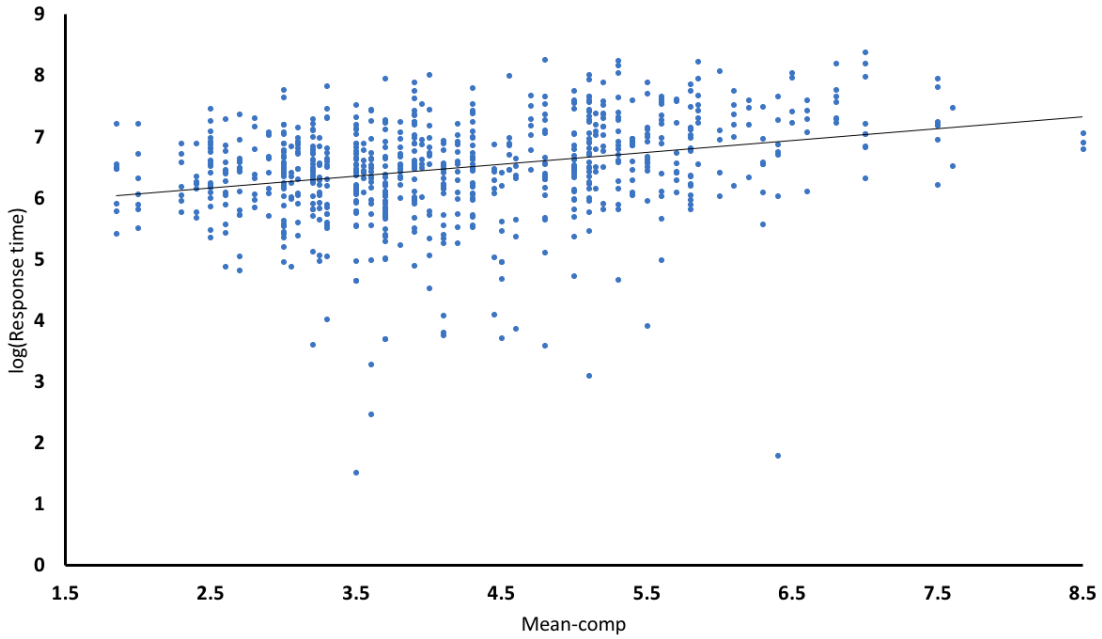


Figure 3.6 Scatterplot of log(response times) in relation to their mean comprehensibility scores

Since there was a significant effect of *Veracity* found in the ANOVA, the emmeans package in *R* was used to generate the estimated mean RTs for true and false statements using the results and coefficients of the lmer model fit. As in Analysis I, it was once again found that false sentences are associated with significantly longer response times (see Table 3.15).

Table 3.15 Back-transformed estimated median response times in ms for both levels of *Veracity*

				95% Confidence Interval	
Sentence	Median RT	SE	df	Lower	Upper
True	520	36.32	36.43	451.45	599.19
False	583	40.68	36.34	506.05	671.56

Discussion

Analysis III examined whether accentedness and comprehensibility ratings for the four voices would accurately predict reaction latencies. Because of the high correlation between the two metrics, only comprehensibility scores were used in the model. This analysis replicated Munro and Derwing's (1995) findings, as results showed that comprehensibility ratings were indeed closely linked to response times, as sentences that had been rated as difficult to understand also required longer to respond to, thereby lending credence to the assumption underlying response time studies that longer RTs reflect increased processing difficulty. This analysis furthermore provided additional evidence that false statements took longer to respond to than true statements.

3.4.4. Analysis IV: Individual voice differences

The analyses presented so far disregarded any talker-specific differences by combining the two Canadian speakers into the 'native-accented' level of *Voice,* while responses to the two Japanese speakers were combined into a single 'foreign-accented' level. To test whether the foreign-accented 'voices elicited significantly different reaction times, the same model used in Analysis I was fit to the dataset, with the two levels for the fixed effect *Voice* now being Foreign1 and Foreign2. Remember that while Foreign1 represented a single speaker, Foreign2 was actually made up of four separate speakers. There were however no overall significant differences between the two 'voices' ($p = 0.262$), as can be seen in Table 3.16 below.

Table 3.16 Type III Analysis of Variance Table with Kenward-Roger's method

	Sum of Squares	Mean Squares	NumDF	DenDF	*F*	*p*
Veracity	0.00	0.00	1	171	0.00	0.998
Voice	0.73	0.73	1	135	1.27	0.262
Face	0.19	0.09	2	169	0.16	0.850
Age	2.94	2.94	1	34	5.07	0.031
Veracity * Voice	2.03	2.03	1	158	3.51	0.063
Veracity * Face	0.19	0.09	2	217	0.16	0.850
Veracity * Age	0.87	0.87	1	157	1.50	0.223
Voice * Face	0.99	0.49	2	173	0.85	0.429
Face * Age	0.07	0.03	2	168	0.06	0.945
Voice * Age	2.07	2.07	1	125	3.57	0.061

Discussion

Based on this analysis, we can conclude that the Foreign1 and Foreign2 voices did not elicit different RTs, which in turn supports the methodological decision in the previous analyses to collapse the two voices into a 'foreign-accented' category, rather than to keep them into the analysis as two unique voices.

3.5. Discussion of Experiment 1

3.5.1. The effect of *Voice*

The experiment's predictions regarding differences in perceived accentedness, intelligibility and comprehensibility between the native and non-native English speakers were borne out in the data, as the experiment successfully replicated Munro and Derwing's (1995) observation that native English listeners take significantly longer to process foreign-accented speech than native-accented speech–despite the Japanese speakers having an overall slower speaking rate than the Canadian speakers. This finding thus appears to confirm that there is a processing cost associated with foreign-accented speech. The accentedness ratings that were obtained in Experiment 1B indicated that the Japanese speakers were given higher ratings of foreign-accentedness than the Canadian speakers, but since the experiment had not included photographs of purported speakers during the accentedness rating task, the data from Experiment 1B provided no insight into the interaction between visual speaker ethnicity and perceived accentedness.

As Munro and Derwing (1995) had found, this study also showed that longer processing times were associated with low comprehensibility ratings, and that listeners were more likely to misidentify foreign-accented sentences as true or false compared to native English speakers' sentences. However, while Munro and Derwing (1995) found no significant effect of sentence veracity, this study's results showed that false statements took longer to process than true statements, which may explain why false statements were also more likely to be correctly identified. Despite these differences in correct identification between native- and foreign-accented speech, the overall estimated probability of correct sentence verification was still high for both native and non-native accents (99% and 94%, respectively). This was once again consistent with Munro and Derwing (1995), as they had reported similarly high accuracy scores for the native and non-native voices (98% and

93%, respectively). Finally, contrary to prediction, participant age was not found to influence either response times or correctness scores.

3.5.2. The effect of *Face*

This experiment's findings are less evident when it comes to the effect of guise ethnicity. As had been discussed in Chapter 2, there were several studies that reported finding evidence of visual speaker ethnicity influencing perceived accentedness ratings, intelligibility, and even speaker evaluations. However, contrary to what had been hypothesized, visual speaker ethnicity did not appear to influence the results, as there were no significant differences in response times between the congruent and incongruent conditions–which would have been in line with exemplar-based models of speech perception–nor was there evidence of listeners performing worse in the Asian face conditions, which would have been consistent with the reverse linguistic stereotyping hypothesis.

3.5.3. Speed/accuracy trade-off

Finally, this experiment showed that that false statements not only took longer to process than true statements, but also that listeners were more likely to correctly identify these false statements compared to the true ones. Taken together, these findings suggest a correlation between the time participants take to respond to a statement and the likelihood of them correctly identifying a sentence's veracity. In other words, there appears to be a trade-off between accuracy and speed.

3.5.4. Limitations

It is tempting to interpret this observed lack of statistical significance for guise ethnicity as evidence that social speaker information does not influence listener performance in semi-online tasks that require an immediate response, but "[c]oncluding from non-significance that there is no effect of an experimental manipulation is a well-known statistical fallacy" (Kirby & Sonderegger, 2018, p. 71). Additionally, the methodological decision to pair all voices with all faces (and to show all these combinations to each listener) may have inadvertently canceled out any effects that guise

ethnicity may have had on RTs or correctness scores. Contrary to other within-subject designs that had kept the face–voice pairings constant throughout, this experiment had exposed participants to all twelve possible face–voice combinations. Post-experiment debriefing all but confirmed the suspicion that this approach had resulted in a disconnect between the guises' face and the voice, as the listeners quickly learned to disregard guise ethnicity entirely. One participant even had to be excluded from the study because he had found the faces to be too distracting and had subsequently closed his eyes for the duration of the experiment. This unintended disassociation between the voices and the faces should serve as a cautionary tale for future experiments. This methodological oversight makes it therefore difficult to draw conclusions from the data regarding the effects of visual speaker information on speech processing. Another important caveat to this experiment is that the sentences spoken by Foreign1 were provided by four different voices, which may have further complicated the interpretation of the effect of *Voice*. The experiments presented in Chapters 4 and 5 were created to address this limitation, and to help determine whether this observed null finding was due to task design, or because visual speaker ethnicity truly does not affect processing.

Chapter 4.

Experiment 2 with videos in Vancouver

The purpose of Experiment 2 is threefold: First, it is a replication of both Munro and Derwing (1995) and Experiment 1, which both found that foreign-accented speech is more difficult to process than native-sounding speech. Second, it serves to test whether the comprehensibility data provides additional evidence for the exemplar theory or the reverse linguistic stereotyping hypothesis. Finally, it aims to further Zheng and Samuel's (2017) investigation into the distinction between interpretation and perception by testing the effect of visual speaker ethnicity on both an online measure (response times) and offline measures of speech evaluation (Likert ratings). This second study was essentially an improved-upon version of Experiment 1, and as such, the study design underwent some significant changes.

First, multi-talker-babble was added to the audio files to prevent the re-occurrence of a ceiling effect in verbal repetition accuracy, and two new tasks were added to provide more in-depth information on the speakers' perceived accentedness and intelligibility. Most importantly, the visual stimulus was changed from a static photograph to a dubbed video to make the guises more realistic. The choice of videos over pictures was also motived by Zheng and Samuel's (2017) observation that videos make the ethnicity manipulations less obvious. which in turn makes participants less likely to 'catch on' to the purpose of the experiment and to change their responses to more socially desirable ones.

While the decision in Experiment 1 to expose participants to all levels of face and voice offered convenience for data analysis purposes, in practice this meant that listeners repeatedly saw the same two faces supposedly 'speaking' with four different voices. An unintended consequence of this choice was that listeners quickly learned to ignore the face altogether. This was problematic, since matched guise studies offer meaningful results only if the listeners buy into the personae that are created by the researcher. To address this, Experiment 2 featured a balanced incomplete block design. The experiment was divided into four blocks, with each block containing 14 sentences (7 true, 7 false) that had been spoken by the same speaker. There were no voice changes within blocks, meaning that each listener saw only one of the four possible face–voice pairings for each

voice. Keeping the face and voice pairings constant throughout the experiment was expected to help convince the listeners that the voices truly belonged to the speakers[15].

To ensure that all sixteen possible guises were shown equally often and in different presentation orders, 64 permutations of the experiment were created. Depending on the assigned group, some participants saw only congruent face–voice pairings (cccc), some saw only incongruent pairings (iiii), and others saw both congruent and incongruent guises (cici/icic). This experiment featured two White Canadian and two Japanese speakers, which meant that there were eight possible congruent *Face–Voice* pairings, and eight incongruent ones. Within the social and cultural context of this experiment (Vancouver), the *Face–Voice* combinations that were hypothesized to be in line with the listeners' expectations were a White speaker with a Canadian accent and an Asian speaker with a Japanese accent. Conversely, White guises speaking with a Japanese accent and ethnically Asian native speakers of English were expected to violate listener expectations. This hypothesis is the same as in Experiment 1 as it is based on Babel and Russell's (2015) Vancouver-based observations, Canadian Census data, and the general expectation that the 'prototypical native speaker' is still more often than not considered to be Caucasian in appearance.

4.1. Methods Experiment 2

4.1.1. Stimuli

This experiment consisted of three different tasks: a sentence verification task, a speech shadowing task, and an accentedness rating task. Most of the sentences that had been used in Experiment 1 were re-used in this sentence verification task. The overall response times to the 56 English sentences in Experiment 1 had revealed that some sentences required noticeably longer processing times, while others were more frequently erroneously identified as correct. For instance, the sentence 'fish live in tall trees' was wrongly identified as correct far more often than other sentences in the dataset. To even out this effect of varying difficulty levels among individual sentences, those with above-

[15] Indeed, during debriefing after the experiment, only a handful of participants indicated that they had noticed something 'unnatural' about the guises, i.e., that they had been manipulated.

average incorrectness scores and unusually long response times were taken out of the dataset and replaced with new ones. These new additions were informally vetted by two Canadian linguists to ensure that they could be easily verified. To create the different face–voice combinations, each speaker's original audio files were paired with the videos of the other three speakers, which resulted in 16 guises (4 speakers x 4 voices). Sixty-four experiment configurations were created to ensure that all sixteen guises were shown in all possible combinations and presentation orders, while an additional sixteen unique configurations were created for the Audio-Only condition.

The speakers were video recorded on an iPad that had been suspended at face-height in a sound-attenuated booth. A high-fidelity microphone simultaneously recorded the audio. The four women each read a list of 64 sentences in its entirety twice. To ensure a natural delivery of the sentences, they were instructed to read the sentence, look up into the camera, and briefly pause before repeating the sentence from memory to ensure that they did not appear to be reading. Speakers were not corrected on their pronunciation except for one instance when a speaker pronounced 'the earth' as 'the ass.' As a result, eight sentences were removed from the stimuli list because mispronunciations had made them unintelligible. This left a total of 56 sentences to be used in this study. See Appendix C for the complete list of sentences.

Due to technical difficulties with the microphone in the sound-attenuated booth with one of the speakers, the audio track that had been captured by the iPad had to be used instead of the high-fidelity recording. Despite the use of a sound-attenuated booth during recording, there was still some background noise that was attenuated in *Audacity* (Audacity Development Team, 2018). The videos were then spliced into sentence-length files and the audio was normalized for peak amplitude. Word-final fricatives were cut at zero crossings, and any noticeable hesitations between words (> 0.5 s) were shortened. Based on the four speakers' multiple renditions of the same sentence, a median duration was calculated for each sentence. The audio files closest to this median value were subsequently selected, after which their length was digitally compressed or lengthened in *Praat* (Boersma & Weenink, 2018) so that all four speakers' renditions of the same sentence were equal in duration.[16] The duration adjustments were made with a 'change duration' command that is part of a plugin called *Praat Vocal Toolkit* (Corretge, 2019).

[16] This was an important change from Experiment 1, where talking speed was unaltered.

Finally, 500 ms silence was added to the beginning of each audio file with another *Praat* script.

The decision to manipulate the speech rate of the four voices was motivated by the observation that the non-native speakers had a noticeably slower speaking rate than the Canadian speakers. Because of this difference in delivery speed, the slower Japanese-accented sentences may have enabled more reliance on prediction than the native-accented ones. By equalizing the duration of the sentences, any potential differences in RTs due to speech rates were minimized. Another reason for creating audio files of equal duration was that it considerably expedited the creation of the stimulus videos. Once a video had been synced to a single audio file, only minor adjustments needed to be made to have the lip movements in that video align with the other three voices' renditions of that same sentence.

To make the task more challenging for the listeners, the decision was made to mix in background noise with the audio to prevent ceiling effects in verbal repetition accuracy (Dragojevic & Giles, 2016). Following McGowan (2015), multi-talker-babble was mixed with the audio files with a signal-to-noise ratio of +5 dB. This SNR was found to be sufficiently challenging without making it impossible to understand the speech in a pilot test that had been administered to two native English speakers without a background in linguistics. The decision to embed the sentences in noise is a departure from the original response time study that this design was based on, as Munro and Derwing (1995) had presented their sentences under ideal listening conditions. Before the multi-talker-babble could be added to the stimulus sentences, remaining differences in amplitude between the four speakers (after normalization) were equalized, as this could cause variability in signal-to-noise ratios (SNRs) across individual files. The LUFs[17], or perceived loudness of the audio files, were adjusted to 89 dB with the help of an *Audacity* plugin called 'ReplayGain,' thus ensuring that all the audio files had roughly the same perceived loudness.

The resulting audio files were placed on a timeline in a video editing program called *Wondershare Filmora* (2019), which allowed the original video to play simultaneously with the edited audio. By cutting frames and by speeding up or slowing down sections of the

[17] LUF = Loudness Units relative to Full scale

video, the two tracks were synced as closely as possible. Because it proved virtually impossible to get all videos and audio files to be perfectly in sync without any obvious imperfections, a feathered mosaic (a blur) was manually placed over the mouth and jaw area in each of the resulting 896 individual videos to cover up any noticeable syncing dysfluencies. This was done in *Adobe Premiere Pro* (see Figure 4.1 below for a visual). The tracking blur automatically adjusted its location and size based on the speakers' head movements and lip and jaw position. Because the blur's opacity allowed large mouth and jaw movements to remain visible (through color changes in the pixels), not all articulatory information was lost, as would have been the case for photographs. It has to be noted that because of this, listeners could still have noticed some instances of misaligned audio-visual information. Screenshots of each of the four speakers without blur can be found in Appendix E. Finally, the ending of all video files had to be edited manually once more in *Quicktime Media Player,* as *Wondershare Filmora* was not a millisecond-specific program, so it had added silences to the end of some of the videos to round off to the nearest .5 s.

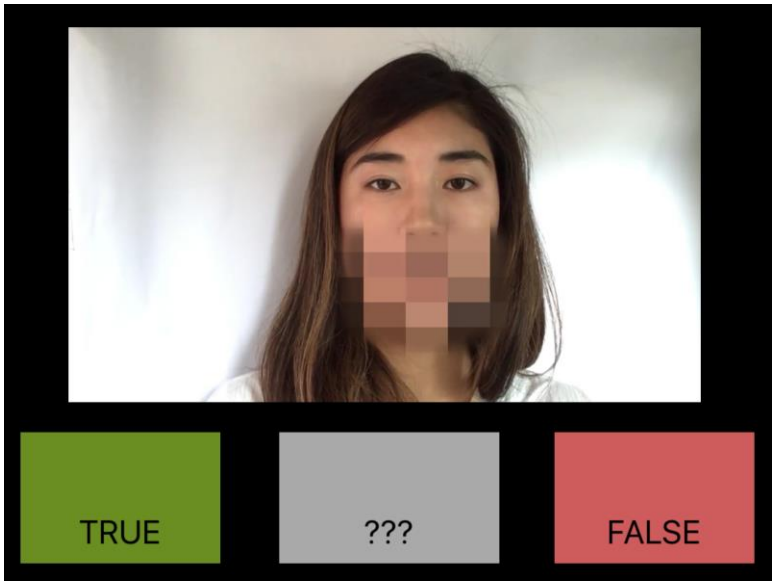


Figure 4.1 A screenshot of what the participants saw during the experiment

4.1.2. Speakers

Two native English-speaking Canadian women and two native Japanese-speaking women lent their voices to this project. The Canadian speakers had both grown up in the same city on Vancouver Island, and were 21 and 22 years old, respectively. The Japanese women had both grown up in Japan and had recently moved to Canada. They were 29

and 23 years old. The researcher herself recorded four practice statements (age 28, slight Dutch accent).

4.1.3. Listeners

Forty-nine listeners were recruited for the Audio-Visual experiment (34 female, 14 male), with an additional thirteen participants for the Audio-Only condition (10 female, 3 male), which served as a control. Assigning only thirteen listeners to the control condition (i.e., 26.5% of the total population) was a purposeful decision. According to White (2018), having a control size between 25% and 30% of the total number of participants "is a good compromise, as this exposes 70% of the sample to the treatment, yet still does not harm power terribly" (Takeaways section, point 2). To get a population sample that was not limited exclusively to students (as is often the case for linguistics studies), most participants were recruited off campus through a public Facebook group where people trade goods and services ('Bunz Vancouver'). A smaller group was recruited on campus– some through direct advertising to departments, others through the Linguistics Department's Research Participation System. The latter group was given course credit, while the other participants received $10 for their time. It is worth noting that only eight participants majored in linguistics; a few of the others had taken between 0 and 4 courses in linguistics, but most were unfamiliar with linguistics and therefore more likely to be naïve to the purpose of the study. All listeners had self-reported normal hearing and were native speakers of English. Thirty-six of the participants were Caucasian, twenty-eight were Asian, and seven were of another ethnicity. As in Experiment 1, there were no restrictions placed on the ethnic backgrounds of the participants. When asked how frequently they interacted with Japanese-accented speakers, fifteen participants said almost never, twenty-nine participants said about once a month, fourteen indicated they interacted with Japanese speakers on a weekly basis, and three said daily. Participants ranged in age between 17 and 58, with a mean age of 27.

4.1.4. Procedure

All participants were tested individually. Depending on the means of recruitment, they were tested in a silent room either on campus or in a public library. Participants were informed that the task they were about to do was a response time study, and that both

speed and accuracy were important. They were explicitly instructed not to move the iPad from its stand and to keep their index finger close to the screen. The researcher (who was present throughout the entire experiment) also told them to keep their eyes focused on the speakers' faces. After signing the informed consent form, they were seated behind an iPad (running on iOS 11.2) that was propped up on a stand in landscape mode. A pair of high-quality noise-cancelling headphones (Model: Sony WH1000XM2/B XB) was connected to the iPad. It had been calibrated to the noise profile in the room to optimize ambient sound cancellation. The volume had been pre-set to a comfortable listening volume. The videos were presented with the *Paradigm for Mobile* app, which is a millisecond accurate stimulus presentation system for iPhone and iPad (Perception Research Systems, 2007).

For task familiarization, the participants first completed four practice statements that featured a video of the experimenter herself. This protocol was intended to convince the participants that the voices they heard in the subsequent experiment also truly belonged to the speakers. The subsequent test items were blocked by speaker, so listeners heard all sixteen sentences in one guise before moving on to the next speaker. What sentences were spoken by which guise was counter-balanced across guises, as was the order in which the speakers were shown. In the Audio-Only condition, the video was replaced by a genderless silhouette. Sentence order within blocks was randomized.

The videos were presented against a black background, along with three large touch response boxes on the bottom of the screen. The left button was green and featured the word TRUE while the right button was red and featured the word FALSE. A third button in the middle read '???'. Participants were instructed to select it when they did not understand what had been said. This third button was added to the experiment to minimize false positives where listeners did not actually understand what had been said but might have guessed correctly anyway. The order of the buttons did not change throughout the experiment. After the video ended, the last frame of the video remained on the screen until a button was selected. This is different from Experiment 1, where the next item would automatically be shown after a period of inactivity. Rating all fifty-six sentences took an average of six minutes.

Once the participants had completed the response time task, they were asked to provide accentedness ratings for each of the four guises on a seven-point scale. This

scale was chosen to make the data more easily comparable to that of Rubin (1992). Participants were randomly shown a still of each of the four speakers without the mosaic on their mouths (so not necessarily in the order of presentation) and were asked to rate the accentedness of that guise on a Likert scale (1 = Canadian accent, 7 = Foreign accent). Since the accentedness ratings were requested only after all four voices had been heard, the participants were forced to recall what voice each guise had spoken with in order to rate their accentedness. In debriefing, most participants admitted that they did not fully remember what each speaker had sounded like. The task's heavy reliance on memory for the AV participants had been deliberate, as I was interested to see whether the listeners' impression and recall of the four speakers' accentedness would possibly be 'colored' by stereotypical default beliefs. In other words, would listeners who had been shown a video of a White woman speaking with a foreign accent 'misremember' that voice as having less of an accent if they had forgotten the voice and were forced to guess? And would listeners misremember the Asian speakers as having *more* of an accent, regardless of what voice that guise had spoken with?

Participants in the Audio-Only condition rated both the voices and the faces separately. This meant that, as in Gnevsheva's (2018) experiment, they were asked to rate the faces on *expected* accentedness, as they had not been given any indication as to which voice supposedly 'belonged' to which face. While the listeners' ratings could be construed as reflective of their stereotypical expectations about what each speaker 'should' sound like, it is important to acknowledge that social desirability bias undoubtedly influenced their ratings.

Although the response time task provided some measure of listener comprehension through its sentence verification scores, additional information on speaker intelligibility was collected at the end of the experiment. During piloting it was discovered that it would be far too onerous to combine the response time task with a separate intelligibility measure, which is why a separate speech shadowing task was added at the end of the experiment. It was meant to provide insight into whether the speakers' ethnicity affected the intelligibility of the four voices in any way. Instead of using the more traditional transcription task, participants were asked to immediately verbally repeat (as opposed to write down) the speech they heard. This type of measure is an improvement over traditional transcription for two reasons: first, it keeps the task in a speaking and listening modality, rather than requiring listeners to switch to the separate system of writing.

Second, it is an immediate measure, which makes it less reliant on memory than transcribing by hand.

Each listener heard the same noise-embedded 28 sentences they had heard earlier, plus 5 randomly chosen sentences in a native accent. The participants' utterances were recorded and transcribed by the researcher afterwards. Because it was relatively easy to check how many mistakes were made, the responses were coded in two different ways; once for number of correct content words, and once for the number of exact word matches. This way the potential influence of different coding approaches could be checked. Once the listeners had completed all three tasks on the iPad, they were given a paper background questionnaire, on which they answered some background questions about their age, their exposure to Japanese-accented speech, how many of their friends spoke with a foreign accent, how many of their friends were of a visible ethnic minority, and what they thought the study had been about.

4.2. Predictions

Since listening to speech under adverse conditions has been predicted to amplify categorical perception (Hanulíková, 2018; Nygaard & Pisoni, 1998), embedding the voice in multi-talker-babble is expected to magnify any possible effects of visual speaker ethnicity on comprehensibility. Since "indexical information complements the processing of linguistic content during a spoken exchange" (McLennan & Luce, 2005), it is furthermore hypothesized that the congruent[18] face–voice combinations will entail faster response times than the incongruent ones. After all, if congruent information complements processing, it would follow that incongruent information should inhibit processing. However, if the Asian guises are found to elicit longer response times regardless of the paired accent, then the data would provide support for the reverse linguistic stereotyping hypothesis, as it posits that listeners will simply 'shut down' when they see a non-Caucasian face (Lippi-Green, 2012).

[18] As a reminder, within the context of this study (listeners in Vancouver), the unexpected pairings are assumed to be a Caucasian guise speaking with a Japanese-accented voice, or an Asian guise speaking with a Canadian English accent. These same local stereotypical expectations were also observed in a Vancouver-based study conducted by Babel and Mellesmoen (2019).

It is expected that the Japanese speakers will be rated as more accented than the Canadian speakers, and that visual speaker ethnicity will modulate these accentedness ratings. If the congruent pairings are rated as less foreign-sounding than their incongruent counterparts, the accentedness data would provide additional evidence for the exemplar theory. Yet if the harshest accentedness ratings are given to the Asian guises regardless of their accent, then the data would be more in line with the reverse linguistic stereotyping hypothesis.

In terms of intelligibility, it is predicted that the foreign-accented voices will elicit less accurate shadowing than the native Canadian voices. If visual speaker ethnicity is also found to influence intelligibility, then exemplar-based models would predict better intelligibility for the same voice in the congruent combinations compared to the incongruent ones, while the reverse linguistic stereotyping hypothesis would predict that the Asian guises will receive lower intelligibility scores, regardless of their accent.

4.3. Results

4.3.1. Analysis I: Response times

This analysis investigated the effects of visual speaker ethnicity, accent, and sentence veracity on reaction times. Observations where the participant either incorrectly identified the veracity of the presented statements (n = 77) or where they had pressed the '???' button (n = 249) were removed from the dataset, which resulted in a disproportionate removal of responses to the Japanese-accented voices (total = 280) as opposed to native-accented sentences (total = 46). Since the model did not attempt to correct this potential downward bias on the mean RTs to Japanese-accented sentences, it should be kept in mind this may potentially have underestimated the differences in response times between the native- and foreign-accented voices. Eight outliers that had studentized residuals of > 4 were removed from the dataset as well.[19] Since the results of the analysis with the

[19] Previous studies have used varying criteria for defining outliers. For instance, Faulkenberry (2017) considered RTs that were "below three and above six median absolute deviations from the overall median" to be "potential contaminant trials" (p. 5), while Marquer and Pereira (1990) removed "RTs that were greater or less than three standard deviations from their respective means" (p. 157). Adank et al. (2009) chose 2.5 standard deviations as their cutoff point, while Kaup, Lüdtke, and Zwaan (2005) chose to remove any responses "longer than 5000 ms or shorter than 200 ms"

outliers included showed no meaningfully different outcomes, they will not be discussed further.

Paradigm recorded each listener's response times in milliseconds, but since it started measuring response times from the moment the stimulus started playing, the audio files' durations had to be subtracted from the total reaction time to calculate a post-sentence reaction time. Besides response times, *Paradigm* also recorded each listener's response (true/false/???) to each sentence and compared this to the correct response. A '???' response and wrong response both received a value of 0, whereas a correct response received a value of 1. Figure 4.2 below shows the spread of RTs with the eight outliers removed. Only the RTs to correct responses were plotted.

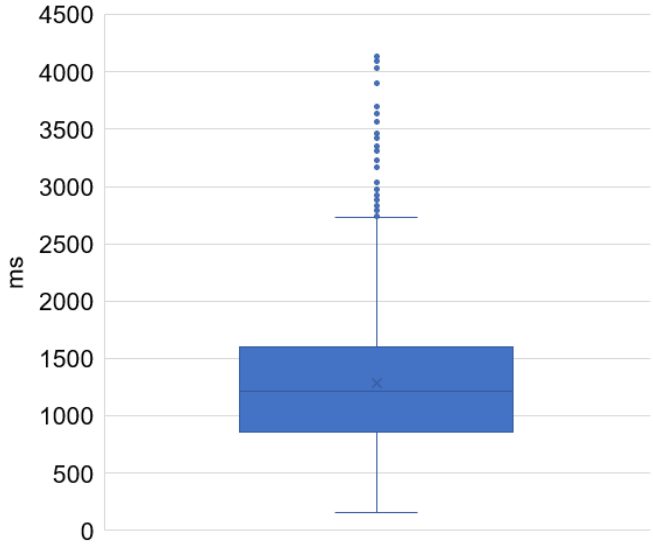


Figure 4.2 Boxplot showing the spread of raw response times in ms

A linear mixed effects model was fitted to the log-transformed response times to examine the relationship between the fixed effects variables *Voice ID* (Foreign1, Foreign2, Native1, Native2), *Face* (Asian, White, no face), *Veracity* (true statement, false statement), and *Trial* (i.e., the number of sentences that each participant was presented with). *Trial* took on values between 1 and 56 and was treated as a numeric variable. The latter

(p. 1117). Kim (2016), on the other hand, rather arbitrarily removed all response times that were shorter than 300 ms and longer than 5000 ms in addition to data points that were three SD removed from the mean.

variable meant to serve as a proxy for the total amount of time a participant had spent answering the questions up to that point to allow testing for learning effects.

In this experiment, the four speakers' voices each constituted a separate level in the fixed effect *Voice ID*, as they were not collapsed into a dichotomous 'accent' variable. Consequently, the results of this study should only be interpreted as pertaining to the specific voices examined, as the findings cannot be extrapolated to all Japanese- and Canadian-accented voices. The speakers' faces, on the other hand, were collapsed into an 'Asian' and 'White' category. This decision to group the four faces into two ethnic categories naturally has its drawbacks (especially considering D'Onofrio's (2019) finding regarding the effect of personas on evaluations), but it was deemed necessary to keep the overall number of comparisons manageable. After all, if there had been 16 face–voice combinations to consider rather than 8, a much higher number of participants would have been required to get sufficient statistical power for each face–voice comparison.

The fitted model also contained the following, normally distributed, random effects: *Sentence* served as a random intercept identifying which sentence a participant was presented with. It helped account for sentence specific effects on response times and allowed the results of the study to be generalized to sentences beyond the 56 used in this study. The random effect *Participant* helped account for any participant-specific main effects on response times, as some participants are inherently faster at responding to stimuli than others. This term also helped account for correlations among participants' multiple responses. Treating this as a random effect also helped generalize the results beyond this study.

Since there is likely to be a correlation between a participant's response times to sentences that had been presented with the same *Face–Voice ID* combination, a random intercept titled *PFV* estimated the interaction effect between *Participant, Face*, and *Voice ID*. This interaction term was coded as a random effect instead of a fixed effect, because we are not actually interested in its estimated values–it simply needed to account for this effect in the model. Treating this as a random effect also helped generalize the results beyond this study. Finally, the random intercept *PFVS* estimated the interaction effect between *Participant, Face, Voice ID,* and Sentence *Veracity*. Similar to *PFV*, each participant had to respond to multiple statements that used the same combination of *Face, Voice ID*, and *Veracity*, so the interaction between these variables can be estimated with

this term. It also helped account for possible correlations between a participant's response times to sentences presented with the same combinations of *Face*, *Voice ID*, and *Veracity*. The reason for not treating this as a fixed effect is the same as that for *PFV*; treating it as a random effect helped to generalize the results beyond this study. Some interactions between these fixed effects were removed in the final model because they were not significant. See Appendix F for a summary of the linear mixed effects model's estimated regression coefficients and the variable coding scheme that was used. The ANOVA table below also shows which interaction terms were included in the final model. As in Experiment 1, the lmerTest package (Bates et al., 2015; Kuznetsova et al., 2017) was used to code dummy variables to produce beta estimates. Note that the data were log-transformed because the assumption of constant variance of the residuals had been violated.

After fitting the above model in *R*, an ANOVA was carried out on this fitted model using the Kenward-Rogers degrees of freedom approximation method with Type 3 Sums of Squares. See the results in Table 4.1.

Table 4.1 Type III Analysis of Variance Table with Kenward-Roger's method

Fixed Effect	Sum of squares	Mean squares	Num DF	Den DF	*F*	*p*
Face	1.11	0.56	2	164	5.24	0.006
Voice ID	8.73	2.91	3	174	27.50	< .001
Veracity	1.09	1.09	1	97	10.33	0.002
Trial	34.24	34.24	1	468	323.55	< .001
Face * Voice ID	1.59	0.27	6	183	2.51	0.023
Face * Veracity	0.05	0.03	2	241	0.24	0.787
Face * Trial	19.61	9.8	2	495	92.61	< .001
Voice ID * Veracity	0.28	0.09	3	238	0.88	0.454
Veracity * Trial	0.51	0.51	1	372	4.86	0.028
Face * Voice ID * Veracity	1.34	0.22	6	234	2.11	0.052

As can be seen in the ANOVA table, there were significant main effects for the four variables examined, in addition to significant 2-way interactions for *Face–Voice ID*, *Face–Trial*, and *Veracity–Trial*. Because there are significant effects involving *Trial*, multiple tests were conducted for specific values of trials, as the difference in means between the various treatment combinations appeared to change over time. This can be seen in Figure

4.3 below. The predicted lines are color coded by levels of *Face*. This plot shows that, over time, response times appeared to steadily increase 4 for the Audio-Visual group, while they remained relatively unchanged for the Audio-Only group.

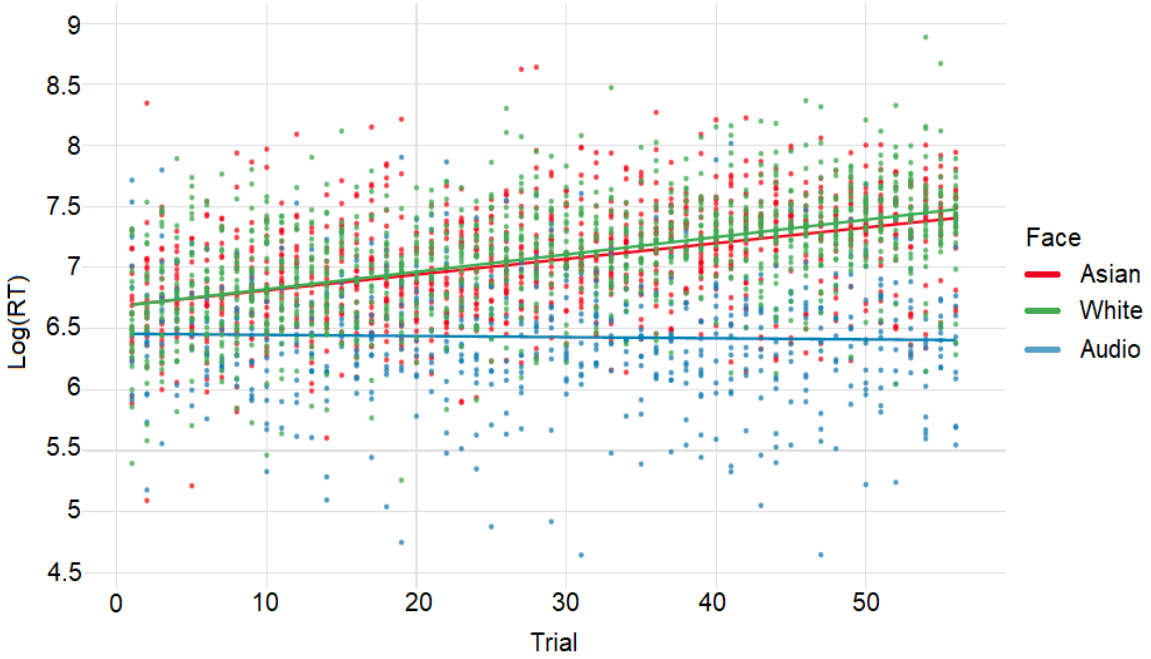


Figure 4.3 Scatterplot of the predicted mean log(Response Time) for each *Face–Voice ID–Veracity* combination relative to *Trial*

This observation was surprising, as it would make more sense for mean response times to decrease as the participants became more familiar with the task. Listener fatigue could not have caused the observed increase in RTs, as we should have seen the same trend in the AO condition. Rather, as will be further discussed in Chapter 5, a more likely explanation for this unexpected trend appears to be a technical difficulty with the stimulus presentation program. I had noticed during data collection that *Paradigm* started to increasingly struggle to retrieve the stimulus videos over Wi-Fi as the experiment progressed. A fitted vs residuals plot was examined to look for evidence of a violation of a linear relationship between (log)RT and covariates. It did not appear to be violated, so we can tentatively assume that the increase in buffering speeds was roughly linear. Nonetheless, this slowly increasing lag in the retrieval and presentation of the stimuli videos may have negatively affected the program's onset errors, which caused the program to register onset times of videos sooner than they were actually presented.

Despite this unexpected trend possibly caused by *Paradigm*, the variable *Trial* was not removed from the analysis as a fixed effect. After all, not including *Trial* and simply averaging over it would cause the mean response times for each *Face–Voice ID* combination to be upwardly biased, as it would ignore the upward trend observed in the data. This would in turn make the Audio-Only vs. Audio-Visual comparisons biased as well, as this trend was not observed in the AO condition. Furthermore, the removal of *Trial* would be allowable only if we were to operate under the assumption that there was either (1) no learning effect, or (2) that there was a learning effect, but that this was overshadowed by the malfunction from the software. If the second assumption is correct, then the effect of *Trial* is not reflective of only a learning effect, but more a kind of average over the combined learning effect and increased RTs due to software malfunction. Since it is impossible to disentangle exactly what caused this upward trend, *Trial* was thus retained in the analysis model.

To address this upwards trend for *Trial*, multiple tests were conducted at specific timepoints to determine the effects of the variables at varying values of trials. It was decided that comparisons between the treatment combinations would be carried out at the 25%, 50%, and 75% quartile values of *Trial*, which corresponds to trials 15, 29, and 43, respectively. Next, interaction plots were generated on the original scale showing the model-estimated RT values at each level of *Face* and *Voice* at the three selected timepoints in the experiment. Figure 4.4 shows these plots:

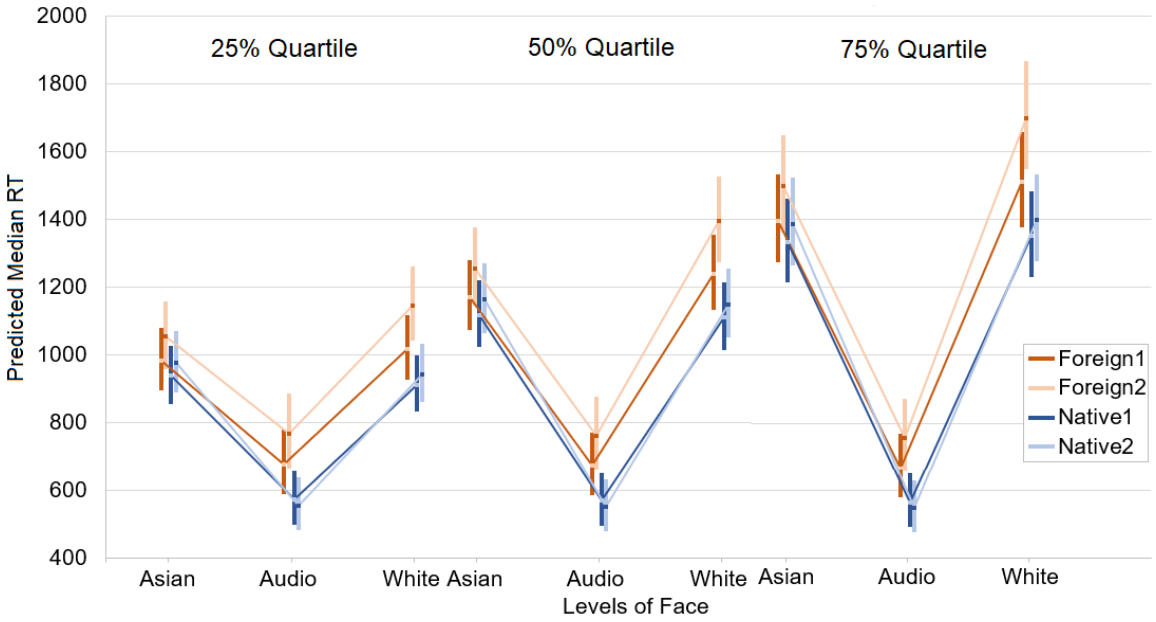


Figure 4.4 Estimated median RTs of each *Face–Voice ID* combination with 95% CIs at the three quartiles. Error bars represent standard errors of the mean

As can be seen, the relative differences in median RTs between the Audio-Visual group and the Audio-Only group increased over time. Overall, it appears that the differences between each *Face–Voice ID* combination stayed fairly consistent, but further analysis was needed to confirm this, especially since there were significant *Face–Voice ID* and *Face–Voice ID–Veracity* interactions. Multiple comparisons were carried out at *Trial* = 15, 29, and 43 using the Tukey-Kramer method to test which pairs of means were significantly different from each other. For the sake of brevity, only the results for *Trial* = 29 are shown below in Table 4.2; the *Face–Voice* comparisons at *Trial* = 15 and 43 are shown in Appendix G. Note that the tables contain the model-predicted medians for each *Face–Voice ID* combination, as the RTs were back-transformed to the original scale. Treatment combinations which share at least one common number in the 'Group' column are not significantly different from one-another. For instance, the Native2–Audio (group 1) pairing is significantly different from all other *Face–Voice* combinations except for the Native1–Audio pairing, whose assigned group value contains 1 (group '12').

Table 4.2 Estimated median response times for each *Face–Voice* combination at *Trial* = 29

Voice ID	Face	Median RT	SE	df	95% Confidence Interval		Group*
					Lower	Upper	
Native2	Audio	552	38.10	109.86	481.04	632.54	1
Native1	Audio	570	39.32	109.43	496.92	653.24	12
Foreign1	Audio	673	46.81	112.78	586.45	772.50	23
Foreign2	Audio	762	54.42	124.99	661.95	878.09	3
Native1	White	1110	50.04	194.90	1015.93	1213.59	4
Native1	Asian	1119	49.80	189.68	1024.91	1221.65	4
Native2	White	1149	51.31	191.07	1052.44	1255.14	4
Native2	Asian	1164	52.71	197.67	1064.52	1272.70	4
Foreign1	Asian	1172	52.66	195.83	1072.80	1280.77	4
Foreign1	White	1241	56.56	203.03	1133.93	1357.26	45
Foreign2	Asian	1257	58.76	224.11	1145.90	1377.82	45
Foreign2	White	1396	64.26	215.87	1274.91	1528.56	5

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

This analysis found no support for either the exemplar theory or the reverse linguistic stereotyping hypothesis, as no clear-cut effects of *Face* (and thus no (in)congruency effects), or *Voice ID* were observed in the Audio-Visual group at any of the three time points. That is to say, there were no significant differences in RTs between the Asian and White face condition for each voice. The only significant differences between voices in the AV condition was between the Foreign2–White face combination (group 5) and the native voices paired with either a White or Asian face (group 4). Additionally, even though voice Foreign2 entailed longer RTs than the two Native voices in the Audio-Only condition, there was no evidence of an overall Foreign vs. Native effect, as Foreign1-Audio was never deemed significantly different from Native1-Audio.

The median RTs in Table 4.2 and Appendix G demonstrate that, on average, Audio-Visual sentences took longer to process than Audio-Only sentences, and that this difference grew larger over time. Yet as was suggested before, the findings in Chapter 5 strongly suggest that this observed difference between the two conditions was most likely caused by equipment failure, rather than indicating some sort of additional processing cost in the AV condition. The latter interpretation would also have run counter to what had been observed in intelligibility research before, where audiovisual input was actually found to *facilitate* intelligibility (Banks et al., 2015).

Two more ANOVAs were carried out on the Audio-Visual data and the Audio-Only data separately to see whether running an ANOVA with fewer comparisons (and therefore increased statistical power) would provide new insights. While no incongruency effects were found in the Audio-Visual comparisons, the Audio-Only analysis did establish a clear *Voice* effect, which is shown below in Table 4.3. Seeing that there was no *Trial* effect, there was no need to do separate analyses at different time points. The subsequent *t*-tests on *Voice ID* confirmed our earlier observation in Experiment 1 and Munro and Derwing's (1995) finding that the non-native voices tended to take longer to process than the native voices, with no differences between the two voices for each accent (see Table 4.4).

Table 4.3 Type III Analysis of Variance Table with Kenward-Roger's method on the AO data

	Sum of squares	Mean squares	NumDF	DenDF	*F*	*p*
Voice ID	7.762	2.587	3	33	15.424	< .0001
Veracity	0.488	0.488	1	55	2.907	0.094
Voice ID * Veracity	0.728	0.243	3	44	1.447	0.242

Table 4.4 Estimated median response times for each level of *Voice ID* in the AO condition

				95% Confidence Interval		
Voice ID	Median RT	SE	df	Lower	Upper	Group*
Native2	552	33.968	34	486.999	625.429	1
Native1	570	34.973	34	502.842	645.420	1
Foreign1	674	42.121	36	593.657	764.977	2
Foreign2	772	50.756	44	675.693	880.926	2

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Despite the above analysis not pointing to an incongruency effect, it was worth formally testing for such an effect using a contrast. Specifically, a hypothesis test was used to determine whether the average estimated median RTs for congruent pairings were significantly different from those of incongruent pairings. Since the contrasts were found to be the same at each of the three values of *Trial*, the table below shows the contrast ratio of just one (*Trial* = 15).

Table 4.5 Contrast of median RTs for congruent and incongruent *Face–Voice* combinations at *Trial* = 15

			95% Confidence Interval			
Contrast ratio	SE	df	Lower	Upper	*t* ratio	*p*
0.955	0.026	205	0.906	1.008	-1.686	0.093

As can be seen in Table 4.5 above, there was no compelling evidence (p = 0.093) that incongruent pairings took longer to respond to than incongruent pairings. Indeed, the estimated median response time to congruent *Face–Voice ID* combinations was 0.955 times that of incongruent pairings.

A final analysis was conducted to formally test whether the addition of a face increased reaction times. As had already become clear in earlier analyses, the Audio-Only treatment combinations tended to have faster response times than Audio-Visual treatment combinations. A contrast was created to formalize whether on average, Audio-Only statements had faster response times than AV statements. This contrast was tested at Trial = 15, 29, and 43. Like the incongruency contrast, converting this contrast back to the original scale means that we are now testing the ratio of median RTs between AV and AO treatment combinations. Below are the results of this contrast at the three values of *Trial*:

Table 4.6 Contrast of median RTs for AV and AO conditions

				95% Confidence Interval			
Trial #	Contrast ratio	SE	df	Lower	Upper	*t* ratio	*p*
15	1.557	0.105	66	1.360	1.782	6.553	< .0001
29	1.89	0.125	60	1.656	2.157	9.647	< .0001
43	2.295	0.155	66	2.006	2.627	12.299	< .0001

At all three values of *Trial*, the tests were significant. At the first quartile (*Trial* = 15), the estimated median response time of Audio-Visual pairings was estimated to be 1.36-1.78 times that of Audio-Only pairings. As the experiment progressed, the median response time of AV pairings increased, with AV pairings taking 1.66-2.16 times longer than AO pairings at Trial = 29 and even double as long at *Trial* = 43 (2.01-2.63). These results show that, overall, Audio-Visual sentences entailed longer RTs than Audio-Only sentences, and that this difference increased as the study progressed.

Finally, a proxy for effect size was generated for the difference in response times between each treatment combination. While traditional significance testing tends to measure differences between population means, the metric used in this study provides a probability measure instead. This metric comes from Browne (2010) as an alternative to Cohen's *d*. The latter could not be calculated because the data had been log-transformed, so the data would not have been on the proper scale. A different metric was therefore used that remains identical on both the original and the log-transformed scale.

Specifically, this metric measures the probability that we would observe a participant in one treatment combination having a slower response time than a participant in another treatment combination. This approach provides more detailed feedback than a simple *p*-value could. For example, a significant difference as determined by a *t*-test could still result in a probability of around 50% that one treatment combination could cause longer RTs than another. Thus, while a *t*-test would label this difference as significant, the actual difference is far less meaningful than previously interpreted. This probability metric was calculated for every possible combination of treatments at *Trial* = 15, 29, and 43. Only the meaningful (i.e., same-voice) comparisons are shown in Table 4.7 below.

Table 4.7 Probability that one *Face–Voice* combination takes longer than another at *Trial* = 15

					95% CI		Probability interval in %	
Voice	Comparison	Ratio*	SE	df	Lower	Upper	Lower	Upper
Foreign1	Asian ÷ White	0.965	0.049	205	0.815	1.143	44.3	53.3
Foreign2	Asian ÷ White	0.919	0.048	226	0.774	1.092	41.8	51.9
Native1	Asian ÷ White	1.029	0.050	183	0.876	1.209	46.7	55.2
Native2	Asian ÷ White	1.034	0.051	190	0.880	1.216	46.8	55.4

* Median RTs in the Asian face condition divided by the Median RTs in the White face condition

The table should be read as follows: the 'Voice' and 'Comparison' columns show what voice and face combinations are compared. The 'Ratio' column contains the ratio of the medians of these two treatment combinations on the original scale. For instance, examining the first row, we can see that the value 0.965 in the 'Ratio' column is the median response time of the Foreign1-Asian combination, divided by the median response time of the Foreign1-White combination. The 'Ratio' column thus tells us that the median response time of the Foreign1-Asian treatment combination was 0.965 times that of the Foreign1-White treatment combination, i.e., the difference was quite small. This

observation is confirmed in the next two columns titled 'Lower CL' and 'Upper CL,' which contain the bounds of a confidence interval for this ratio. They show that the estimated median response time of Foreign1-Asian was between 0.815-1.143 times that of Foreign1-White. This confidence interval contains 1, indicating that the difference between these two treatments was not significant. The last two columns contain the confidence interval for the probability in percentages that a randomly chosen participant reacting to sentences in the (congruent) Foreign1-Asian treatment combination will have a shorter response time than another participant responding to sentences in the (incongruent) Foreign1-White treatment combination. In this example, the probability interval is between 44% and 53%. Looking at the data in Table 4.7, then, it becomes clear that there are no significant differences between the different face conditions, suggesting that any observed differences between the two conditions are coincidental rather than systematic.

Table 4.6 had shown that median response times were significantly longer in the AV condition compared to the AO condition, so we would expect the probability that someone in the AV condition would take longer to respond than someone in an AO condition to be over 50%. Table 4.8 below does indeed show that the Audio-Visual condition entailed longer response times than the Audio-Only condition for all four voices.

Table 4.8 Probability that the AV condition takes longer than the AO condition at *Trial* = 15

					95% CI		Probability interval in %	
Voice	Comparison	Ratio	SE	df	Lower	Upper	Lower	Upper
Foreign1	Asian ÷ Audio	1.449	0.117	123	1.107	1.897	56.8	67
Foreign1	White ÷ Audio	1.502	0.054	123	1.147	1.967	58.1	68.4
Foreign2	Asian ÷ Audio	1.372	0.115	140	1.039	1.811	55.5	67.4
Foreign2	White ÷ Audio	1.492	0.056	138	1.131	1.968	58.4	69.9
Native1	Asian ÷ Audio	1.635	0.130	117	1.253	2.133	60.4	70
Native1	White ÷ Audio	1.588	0.050	118	1.217	2.073	59.7	69.5
Native2	Asian ÷ Audio	1.756	0.141	121	1.343	2.297	62.8	72.4
Native2	White ÷ Audio	1.698	0.047	119	1.301	2.217	61.7	71.3

The probability ratios for *Trial* = 29, 43 are included in Appendix H. They are similar to the tables discussed above, with none of them providing evidence of an incongruency effect. The differences in RTs between the AV and AO conditions do increase as the experiment progresses, which can be seen in the increased probability intervals.

Discussion Analysis I

This analysis yielded no evidence of visual speaker ethnicity affecting comprehensibility, as there were no significant differences in RTs between the different levels of *Face*. This analysis thus fails to provide support for either of the two competing theories described in section 4.2, as both the exemplar theory framework and the reverse linguistic stereotyping hypothesis can be understood to presuppose an effect of visual speaker ethnicity. The present finding that processing speeds appear to be unaffected by this type of visual social information seems to be consistent with Squires (2013), who also failed to find an effect of her speakers' visible social information on response times.

This study does corroborate previous findings that there is a processing cost associated with unfamiliar foreign-accented speech, as participants responded significantly slower to the Japanese-accented voices compared to the familiar Canadian-accented ones–in spite of the fact that the model may have underestimated these differences due to the fact that many more response times to the foreign-accented voices had to be removed from the model because they had been misidentified. This finding is consistent with previous research (Floccia et al., 2006; Munro & Derwing, 1995; Perry et al., 2018), which also reported that unfamiliar regional or foreign accents take longer to process. Finally, although participants in the Audio-Visual condition required more time to respond than the participants in the Audio-Visual treatment group, this difference between the AV and AO condition may be due to a malfunction of the stimulus presentation program and should therefore not be interpreted as evidence of there being an additional processing cost for Audio-Visual input.

4.3.2. Analysis II: Accentedness ratings

The next analysis investigated whether accentedness ratings differed as a function of face, i.e., whether the same voice received different accentedness ratings depending on what face it had been paired with. Remember that for this task, each participant provided one accentedness rating for each of the four guises they had been exposed to, so four in total. Since there was a roughly even spread of accentedness scores, the ratings could be treated as a continuous variable and a linear mixed effects model was fitted to the data. The model contained the fixed effects *Voice ID* and *Face*, and the normally distributed random effect *Participant.* Table 4.9 below shows the ANOVA results using

type-3 sums of squares and the Kenward-Rogers degree-of-freedom approximation. Both main effects were significant, including a significant interaction effect between *Face* and *Voice ID*.

Table 4.9 Type III Analysis of Variance Table with Kenward-Roger's method

Fixed Effect	Sum of squares	Mean squares	NumDF	DenDF	*F*	*p*
Face	23.57	11.78	2	117	6.28	0.003
Voice ID	633.76	211.26	3	176	113.23	< .001
Face * Voice ID	30.58	5.10	6	211	2.726	0.014

The interaction plot in Figure 4.5 below illustrates the main effect of *Voice ID*, as the Japanese speakers were given noticeably higher accentedness ratings than the Canadian speakers, but it does not show an obvious effect of *Face* or an interaction effect between the two variables. Looking at the interaction plot, it does appear that each voice received slightly lower overall accentedness ratings in the congruent condition than in the incongruent condition. As a reminder: 1 = Canadian-sounding, and 7 = 'foreign-sounding.'

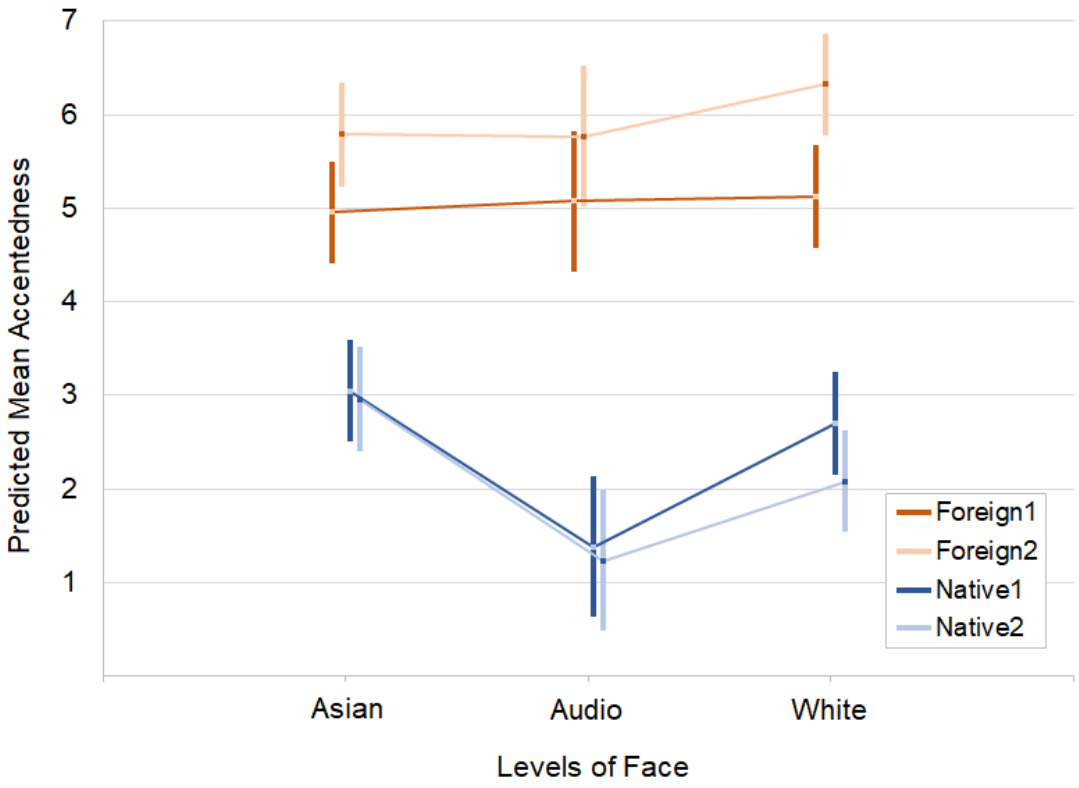


Figure 4.5 Interaction plot of estimated mean accentedness scores for each *Face–Voice* combination. Error bars represent standard errors of the mean

To further probe the interaction between *Face* and *Voice ID*, a Tukey-Kramer test was conducted to test which pairs of model-estimated means were significantly different from each other. The results are presented in Table 4.10, and Figure 4.6.

Table 4.10 Model-estimated mean accentedness ratings for each *Face–Voice* combination

					95% Confidence Interval		
Voice ID	Face	Mean	SE	df	Lower	Upper	Group*
Native2	Audio	1.231	0.381	236	0.481	1.980	1
Native1	Audio	1.385	0.381	236	0.635	2.134	1
Native2	White	2.079	0.275	236	1.537	2.621	12
Native1	White	2.704	0.281	236	2.150	3.257	12
Native2	Asian	2.960	0.281	236	2.406	3.513	2
Native1	Asian	3.044	0.275	236	2.502	3.587	2
Foreign1	Asian	4.959	0.275	236	4.417	5.501	3
Foreign1	Audio	5.077	0.381	236	4.327	5.827	34
Foreign1	White	5.126	0.281	236	4.573	5.680	34
Foreign2	Audio	5.769	0.381	236	5.020	6.519	34
Foreign2	Asian	5.787	0.281	236	5.234	6.340	34
Foreign2	White	6.324	0.275	236	5.782	6.867	4

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

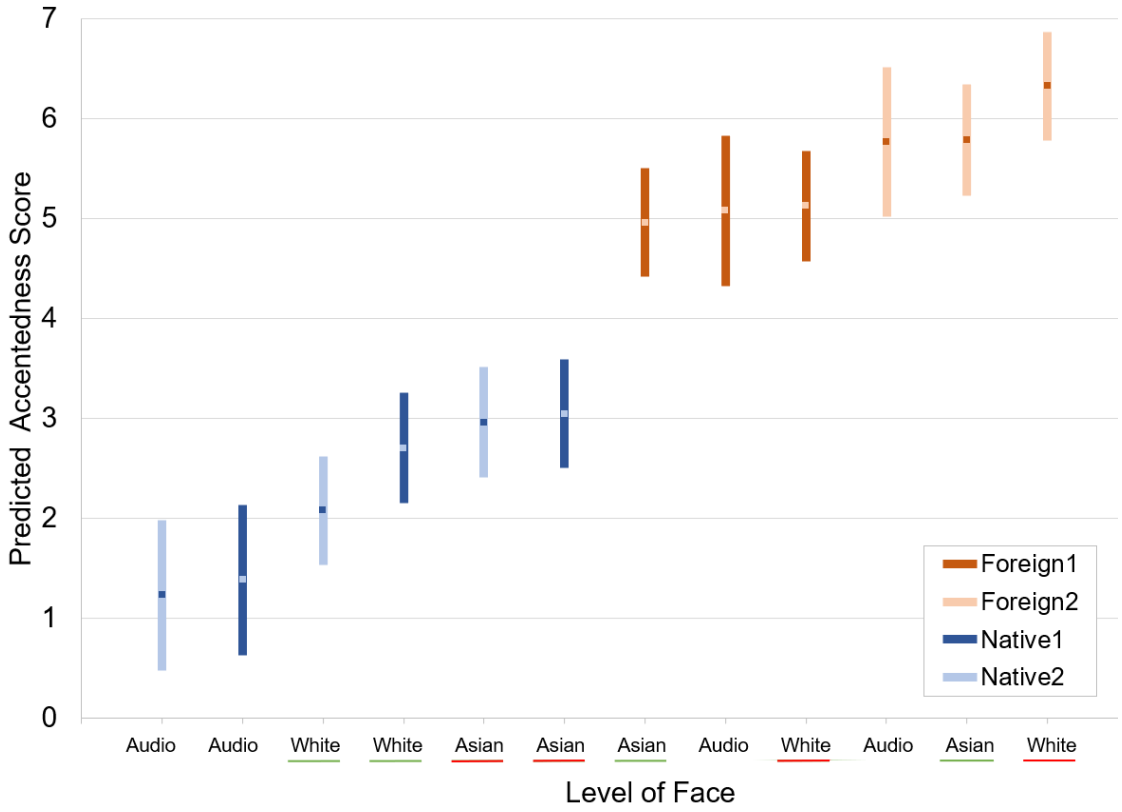


Figure 4.6 Estimated mean accentedness ratings for each *Face–Voice* combination with SE bars

Yet contrary to what was hypothesized, the accentedness ratings appeared to be unaffected by guise ethnicity, as there were no significant differences between the Asian and White conditions for each voice. The foreign-accented voices did receive higher overall accentedness ratings than the native voices, as predicted. Also, the two native voices received significantly lower accentedness ratings in the Audio-Only condition (assigned to 'Group 1') compared to the incongruent Asian face condition (assigned to 'Group 2'), which hints at the possibility that the perceived accentedness of native voices alone may be negatively impacted by the *addition* of an unexpected face. However, since there were no differences between the congruent and incongruent Audio-Visual conditions, this data provided no support for either the exemplar theory or the RLS hypothesis.

Discussion Analysis II

This analysis supported the prediction that the two Japanese-accented voices would receive higher accentedness ratings than the two Canadian-accented voices. Since

no effect of visual speaker ethnicity was observed, there was also no evidence of an (in)congruency effect or racial bias influencing accentedness ratings. While this finding runs counter to findings in some of the existing literature (e.g., Babel & Russell, 2015; Hanulíková, 2018; Kang & Rubin, 2009; McGowan, 2015; Rubin, 1992), this observed lack of evidence for visual speaker ethnicity affecting perceived accentedness has in fact been reported in other, more recent experiments as well (see Hanulíková, 2018; Karpinska, 2019; McCrocklin et al., 2018).

The contradictory findings across different studies lend credence to Zheng and Samuel's (2017) suggestion that methodological choices in perceptual studies can yield different results, as accent judgments appear to be highly sensitive to contextual factors and methodological choices. Zheng and Samuel (2017) were only able to replicate Rubin's observed effects of visual speaker ethnicity on accent judgments when they used a similar blocked between-subjects design, but this effect disappeared when they changed the methodology. When they presented the foreign-accented and native-sounding guises in a mixed rather than a blocked design, they did not find an effect of visual speaker ethnicity on accentedness. These findings raise the possibility that the previously reported effects of visual speaker ethnicity on accentedness ratings are attributable to methodological artefacts.

4.3.3. Analysis III: Intelligibility scores

The experiment's third and final analysis looked at the effect of visual speaker ethnicity and accent on intelligibility scores. As was discussed in the procedure section, participants were asked to orally repeat the sentences they heard, rather than write them down. Their recorded responses were transcribed and scored in two ways: once for exact word matches, and once for correct content words only. Since the scores were very similar for both approaches, the decision was made to analyze only the percentage of correct content words. All subsequent modelling was therefore done using content score data only.

Most participants perfectly repeated most sentences, which can partially be attributed to the fact that the participants had already heard the sentences once before in the response time task. This high number of perfect content scores made it difficult fitting a model to the data. As a result, the content scores were collapsed into either '100 percent

correct,' or 'less than 100 percent correct' for each sentence.[20] Below is a table showing the breakdown of this new metric for each *Voice ID*.

Table 4.11 Number of (<)100% correctly understood sentences

Voice ID	<100% correct	100% correct
Foreign1	63	805
Foreign2	212	656
Native1	1	151
Native2	4	154

Because of the extremely low number of native sentences that were not correctly repeated (n = 5), the native voices were removed from the model. Because the content scores were collapsed into a 2-level factor variable (perfect/imperfect content score), I opted to model the *probability* of an individual achieving a perfect content score. A mixed logistic regression model was fitted to the data. The model had the following fixed effects: *Voice ID* (Foreign1, Foreign2)*, Face* (White, Asian, Audio-Only)*, Veracity* (True, False)*,* and *Trial* (1-56). It also included all six possible two-way interactions between these fixed effects.

The mixed logistic regression had *Sentence* and *Participant* as random effects. The *PFV* and *PFVA* random effects were not included in this analysis because they made the model too complicated to be fit adequately. Below is the Type-3 sum of squares ANOVA that was carried out on the fitted model using Kenward-Rogers for the degrees of freedom. It shows that only *Voice ID* has a significant main effect, and that there was no significant interaction between *Face* and *Voice ID*. Yet because both the *Voice ID * Veracity* and the *Face * Veracity* interactions are significant, *Face* and *Veracity* both still play a role in the model.

[20] This approach to calculating verbal repetition accuracy was far simpler than that of Babel and Russell (2015), who normalized "the proportion correct…to rationalized arcsine units (RAUs)" and then used the RAUs "as the dependent measure in a hierarchical linear regression model with RAU scores calculated per sentence (p. 2827).

Table 4.12 Type III Tests of Fixed Effects

Effect	Num DF	Den DF	*F*	Pr > *F*
Face	2	878	0.38	0.683
Voice ID	1	1721	21.59	< .001
Veracity	1	92	0.24	0.629
Trial	1	1721	1.28	0.259
Face * Voice ID	2	1721	0.04	0.959
Trial * Face	2	1721	0.43	0.652
Face * Veracity	2	1721	3.58	0.028
Trial * Voice ID	1	1721	0.62	0.431
Voice ID * Veracity	1	1721	8.66	0.003
Trial * Veracity	1	1721	0.40	0.527

As there was no significant interaction between *Voice ID* and *Face*, and no significant main effect for *Face* or *Veracity*, we only need to look at the model-estimated probabilities of perfect verbal repetition for each *Voice ID*, which is found in Table 4.13.

Table 4.13 Estimated probability of listeners correctly repeating the two accented voices

			95% Confidence Interval		
Voice ID	Probability	SE	Lower	Upper	Group*
Foreign1	0.971	0.008	0.949	0.984	1
Foreign2	0.823	0.037	0.736	0.886	2

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

When averaged over *Veracity* and *Face*, the probability of a participant perfectly repeating a sentence spoken by Foreign1 is 97%, while this probability is significantly lower (82%) for Foreign2. In other words, listeners had more difficulty understanding (and consequently repeating) Foreign2 compared to Foreign1. This finding is also illustrated in the interaction plot shown in Figure 4.7 below, showing far greater between-subject variability for Foreign2 than for Foreign1.

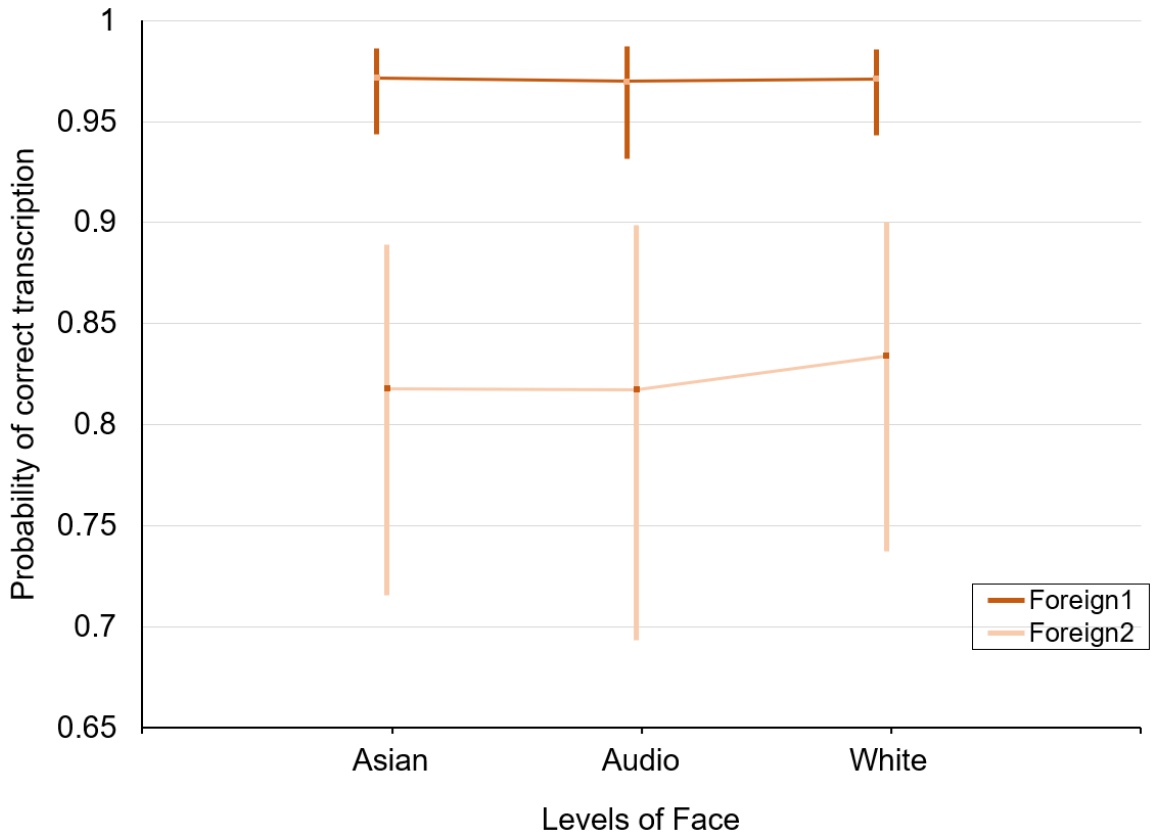


Figure 4.7 **Interaction plot of the probabilities of perfect repetition for the two foreign-accented voices in each *Face* combination. Error bars represent standard errors of the mean**

Given the significant interaction between *Voice ID* and *Veracity*, it was worthwhile to look at a breakdown of the two voices' respective probabilities of correct repetition for each sentence type (i.e., true or false sentences). As can be seen in Table 4.14, there appears to be some variation between perfect repetition probabilities depending on the sentence veracity within Foreign2, but these were not significantly different.

Table 4.14 **Correct repetition probabilities for each voice at the two levels of *Veracity***

				95% Confidence Interval		
Voice ID	Veracity	Probability	SE	Lower	Upper	Group*
Foreign1	False	0.960	0.015	0.916	0.981	1
Foreign1	True	0.979	0.009	0.952	0.991	1
Foreign2	False	0.857	0.044	0.745	0.925	2
Foreign2	True	0.784	0.060	0.642	0.880	2

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Discussion Analysis III

This third and final analysis investigated whether the voices' intelligibility scores differed as a function of what face they had been paired with. Results did not support the exemplar theory, as congruent face–voice combinations did not receive higher repetition scores than incongruent ones, nor did they support the reverse linguistic stereotyping hypothesis, as visual speaker ethnicity did not appear to affect intelligibility scores whatsoever. There was however clear evidence that the native voices were more intelligible than the foreign voices, and that Foreign2 was less likely to be correctly repeated than Foreign1.

4.4. Discussion of Experiment 2

This experiment was designed to explore whether native and non-native speakers of different ethnicities would elicit different response times, accentedness ratings, and intelligibility scores. In addition to testing Munro and Derwing's (1995) finding that foreign-accented speakers are more difficult to process than native-sounding speakers, this experiment also tested the competing predictions of the exemplar theory and the reverse linguistic stereotyping hypothesis with regards to the guises' intelligibility scores. By adding visual speaker information to each speech sample, the potential influence of visual speaker ethnicity and face–voice (in)congruency was also tested on response latencies and accentedness ratings.

Results revealed that contrary to what had been hypothesized, embedding the voices in noise did not magnify the predicted effect of visual speaker ethnicity, as there appeared to be no evidence of such an ethnicity effect on (1) response times, (2) intelligibility, or (3) accentedness ratings. As such, the outcomes fail to provide support for the exemplar theory and the reverse linguistic stereotyping hypothesis, as both theoretical frameworks presuppose an effect of social speaker information. Like Experiment 1, this experiment did however replicate Munro and Derwing's (1995) finding that foreign-accented voices have longer response times (interpreted as lower comprehensibility) than native-accented voices, as responses to the two Canadian-accented voices were indeed faster than those to the two Japanese-accented voices. The Canadian voices also received overall lower accentedness ratings and were more likely to be correctly repeated compared to their Japanese-accented counterparts.

Previous research findings on the effect of visual speaker ethnicity on intelligibility and accentedness ratings have been inconsistent. While a fair number of researchers have reported such an effect (e.g., Babel & Russell, 2015; Hanulíková, 2018; Kang & Rubin, 2009; McGowan, 2015; Rubin, 1992), this research and a few others (see for instance Karpinska, 2019; McCrocklin et al., 2018) did not. Contradictory findings have even been reported within the same study; Gnevsheva (2018), for instance, found an expectation mismatch effect for White German speakers, but not for Korean speakers. These seemingly opposing findings thus hint at the possibility that the effect may be dependent on researchers' methodological choices and their interpretation of the results. The fact that many of the discussed studies did not report effect sizes in their analysis (i.e., Babel & Russell, 2015; Gnevsheva, 2018; Hanulíková, 2018; Karpinska, 2019; McGowan, 2015; Yi, Smiljanic, & Chandrasekaran, 2014) has made it more difficult to compare the different studies' findings and to determine whether their observed effects were even large enough to be meaningful. After all, a significant difference with a small effect size would still be considered inconsequential. A good example of a study where even without effect sizes, the results seem more of a coincidence rather than evidence of a systematic effect is Yi et al. (2013), who concluded that a 0.9% difference in accentedness ratings between an AV and AO condition evidenced listener bias, because their statistical analysis had labeled it as significant. Findings such as these cannot be accepted without replication and other further probing.

This experiment used similar offline tasks to those that did yield an effect of visual speaker ethnicity, but some notable differences in approach may have contributed to the diverging results reported for the offline tasks. Firstly, following Zheng and Samuel (2017), the decision to use realistic videos rather than photographs was expected to help reduce the demand characteristics, as the guise ethnicity manipulation was assumed to be less obvious than using photographs. This in turn would make it less likely that listeners would alter their responses based on what they believed the experiment was investigating. This observation had been made by Zheng and Samuel (2017), who noticed in their series of experiments that the observed effect of visual speaker ethnicity on accentedness largely disappeared when they used videos instead of photographs. Secondly, the accentedness ratings at the very end of the experiment were heavily reliant on listeners' ability to recall what each guise had sounded like. This approach may have made the ratings more susceptible to social desirability bias, which may in turn have influenced the listeners

ratings (but *not* their perception). Thirdly, the intelligibility task was done verbally, which meant that listeners were able to respond much faster than if they had to type it out–during which they would run the risk of making spelling mistakes or forgetting words because of the time it would take to type.

This experiment was the first to use a semi-online measure to measure listeners' reaction latencies to various guises. The fact that visual speaker ethnicity did not appear to affect comprehensibility provides some tentative support for the suggestion that speech processing may be unaffected by social speaker information. However, based on the results from this study, we cannot assume that it is the offline or online nature of the task that caused different results, seeing as visual speaker ethnicity was not found in either offline or online tasks. As such, it is not yet possible to confirm the premise that speech perception may be entirely separate from post-stimulus interpretation (see: Firestone & Scholl, 2016; Zheng & Samuel, 2017).

Chapter 5.

Experiment 3 with videos in Amsterdam

So far, most studies investigating the effect of visual speaker ethnicity on how speech is interpreted and perceived have focused on native English listeners. To make the results less Anglo-centric and less tied to the Vancouver context, the third experiment in this series was conducted on native Dutch listeners in the Netherlands. Following Grondelaers, van Gent and van Hout (2015), data collection took place in two Dutch cities: Amsterdam and Nijmegen.

The largest immigrant groups in the Vancouver Metropolitan Area are from China, the Philippines, and Hong Kong (Statistics Canada, 2017a), while the two largest immigrant groups in Amsterdam hail from the former Dutch colonies Suriname and the Antilles. Yet the Dutch tend not to perceive the Surinamese and the Antilleans as 'the prototypical immigrant.' Instead, it is the Moroccans and the Turks who are the biggest two immigrant groups in the Netherlands as a whole (CBS, 2018, p. 29). In Amsterdam, however, they are the respective second and third-largest immigrant groups (OIS Amsterdam, 2018), while they rank as the fourth and second-largest immigrant groups in Nijmegen (Centraal Bureau voor de Statistiek, 2019).

Note that these immigrant groups tend to favor the urban centers over the rural areas; the country's four biggest cities (i.e., Amsterdam, Rotterdam, the Hague, and Utrecht) together house almost half of *all* the Surinamese and Moroccan immigrants, whereas only nine percent of people with Dutch heritage live in one of those cities (Centraal Bureau voor de Statistiek, 2018, p. 41). City-dwellers are therefore far more likely to encounter immigrants in their daily lives than people living in the more rural areas of the Netherlands.

5.1. Opinions on immigration in Canada vs. the Netherlands

Canadian and Dutch attitudes toward assimilation are generally characterized as different. Where the Dutch tend to prefer immigrants to assimilate with the local population (i.e., only the host culture is important), Canadians are usually assumed to favor

integration instead–a type of adaptation where immigrants are encouraged to maintain some aspects their culture while blending it with their new national identity (Arends-Tóth & Vijver, 2003; Kymlicka, 2012; Mahtani & Mountz, 2002; van Oudenhoven, Prins, & Buunk, 1998; van Oudenhoven, Ward, & Masgoret, 2006). It has been suggested that the general belief among Dutch nationals is that the Turkish and Moroccan immigrants consider their original culture as more important than the host culture, and that they favor separation over integration or assimilation (van Oudenhoven et al., 1998).

The Dutch are also said to perceive Moroccan and Turkish immigrants as "the two groups lowest in the ethnic hierarchy…with the weakest socio-economic position in Dutch society and the greatest discrepancy between their cultural practices and the Dutch way of life" (Hagendoorn, Veenman, & Vollebergh, 2003, p. 10). There is also a tendency to associate young Moroccan men with crime. Social statistics do suggest that this stereotype is at least partially rooted in facts[21], as first- and second-generation Moroccans between the ages of 18 and 25 are indeed almost five times more likely to face criminal charges than Dutch youth of the same age (Centraal Bureau voor de Statistiek, 2018, p. 105).

The speech variety associated with Moroccan immigrants has been termed 'Moroccan-flavored Dutch' (MFD)[22] and is seen as a low-prestige, non-standard ethnolect (Grondelaers et al., 2015; Nortier & Dorleijn, 2008). MFD is predominantly spoken by young people of Moroccan descent, but not exclusively so; the vernacular has covert prestige among young people of other ethnic origins as well, including those with a Dutch ethnic background (Dorleijn & Nortier, 2006). The variety has even started to take on the status of a general 'ethnic' or 'immigrant' accent: "standaardallochtoons" [Standard Immigrant Language] (Klok, 2008). The variety can evoke a host of negative associations among fellow Dutch nationals. This stigma became clear in a free-response study conducted on 172 highly educated Dutch participants; when asked what descriptors first came to mind when thinking of "Dutch with a Moroccan accent," some of the most

[21] Whether or not these statistics are the result of racial profiling by the police is beyond the remit of this dissertation.

[22] See Nortier and Dorleijn (2008) for an introduction to the features of MFD and its sociolinguistic position in Dutch society.

frequently-mentioned adjectives were 'aggressive,' 'unadapted,' 'foreign,' 'lowly educated,' 'anti-social,' but also 'unintelligible' and 'ugly' (Grondelaers & Speelman, 2015).

The findings of a recent speech evaluation experiment paint a slightly less grim picture, as Grondelaers and van Gent (2019) reported that MFD was assigned higher dynamism ratings[23] than the prestigious Standard Dutch (SD) variety and a regionally accented variety. Yet when it came to superiority ratings[24] MFD speakers still received far lower scores than the other two varieties, with those who had a strong accent receiving even lower scores than those who had a mild accent. Interestingly, superiority ratings were also affected by indexical information, as the mere suggestion that a speaker had a Moroccan background (by giving him an Arabic name) caused a downgrade in superiority ratings *even when* the speaker had no discernable ethnic accent (Grondelaers & van Gent, 2019).

These findings are not surprising within the Dutch context. Over the last few decades, there has been a growing anti-immigration sentiment in some Western countries, with Muslims specifically bearing the brunt. While the Netherlands–like so many other Western countries–seems to at least partially share this xenophobic sentiment, Canada continues to have one of the most favorable public attitudes towards immigration in the world (Grondelaers et al., 2015; Hiebert, 2006). A recent survey of public perceptions towards immigration shows that Canadians generally hold positive views towards immigration, but that they do have growing concerns about the numbers of immigrants entering the country on a yearly basis.[25] When it comes to Canadian views on Asia in particular, another survey found that 74% of Canadians believe that the economy is positively impacted by immigration from Asia (The Asia Pacific Foundation of Canada, 2018, p. 8). It should be pointed out that public opinion in Canada is highly regionalized, with noticeable differences between people who live in more rural areas compared to those who live in large urban centers–which is where most immigrants tend to settle (Angus Reid Institute, 2018; Mahtani & Mountz, 2002).

---

[23] Operationalized by the following Likert- statements: *This person… is [modern] [cool] [hip] [tough].*

[24] Operationalized by the following Likert- statements: *This person…[is posh] [has got good grades] [has good management skills] [has made a lot of money] [has a lot of professional experience].*

[25] It merits mention that these findings were mostly based on the question "Would you say 310,000 new permanent residents in 2018 is [too many] [about right] [too few] [not sure]," without telling participants that this number equates to only 0.8 per cent of the total population.

In 2019, the Pew Research Center published the findings of a survey which had examined public attitudes towards immigration in eighteen countries that together host more than half of the world's migrants. Canadians were ranked as the most positively inclined towards immigrants of all eighteen countries, with only 27% of Canadians considering immigrants a burden (rather than a strength). This stands in sharp contrast with the Netherlands, which was ranked 13th because 42% of its population had indicated seeing immigrants as a burden. Another 42% of Dutch respondents agreed that immigrants were more to blame for crime than other groups, compared to only 17% of Canadians. For both countries, the people who were more likely to view immigrants as a strength were 1) those on the ideological left, 2) young people, 3) highly-educated individuals, or 4) high earners (Gonzalez-Barrera & Connor, 2019).

5.2. Study goals

This response time study will contrast the most prestigious standard native variety of Dutch with Moroccan-flavored Dutch, and as such, will provide valuable data on the evaluation and processing of an ethnolect that is understudied. Whilst the available literature on MFD has mostly focused on describing the variety's features and evaluative judgments of its speakers, there is yet to be a study surveying the comprehensibility of MFD. The effect of a visible Moroccan identity on the perception of native Dutch has been investigated only by Hanulíková (2018), who was unable to find an effect of visual speaker ethnicity on recall or comprehensibility. She did note higher accentedness ratings for the Moroccan guises under adverse listening conditions. As was theorized in Section 2.2.1, this observed increased accentedness score may have been due a misattribution of processing difficulty. To further investigate this, the following experiment will examine the effect of ethnicity (in this case, White vs. Moroccan), on the comprehensibility, accentedness, and perceived credibility of both SD and MFD under adverse listening conditions.

By using the voices of female speakers of MFD, this experiment will provide valuable data on how the speech of female speakers of MFD is processed and evaluated, as data on this particular group is non-existent. Women do not seem to be a popular choice in ethnolinguistic studies focusing on MFD, as the majority of these studies have been conducted on male speakers exclusively (e.g., Cornips, 2002; Grondelaers & van Gent,

2019; Grondelaers et al., 2015; Jaspers, 2004, 2005). Nortier's (2017) article looking at "metalinguistic comments on women and girls using youth language in the Moroccan community in the Netherlands" (p. 15) appears to be the only research to date that specifically focuses on MFD in relation to women.

Seeing that research into MFD is still relatively limited, this preference for male speakers is understandable. After all, dialectal research has shown us that it is men who tend to favor varieties with covert prestige, whereas women generally favor socially prestigious varieties. This makes it less likely to find strong ethnolectal features in the speech of female MFD speakers. Because of this, they are not seen as 'typical speakers' of MFD, which in turn begs the question whether some of the negative adjectives (i.e., 'aggressive,' 'foreign,' 'anti-social,' and 'unintelligible') that were documented by Grondelaers and Speelman (2015) would also be applicable to female speakers of MFD, as participants may have been thinking about the prototypical male speaker when supplying these descriptors.

One of the reasons for investigating this phenomenon in a cross-cultural context is to examine the extent to which varying attitudes towards specific immigrant groups can influence speech evaluation and processing differently. Because the two countries have different immigrant populations, Vancouverites are more likely to imagine the prototypical immigrant as 'Asian' (possibly along with all the associated stereotypes of that group), while the Dutch are more likely to think of a 'Middle Eastern' face instead. Knowing that the Dutch seem to hold more negative attitudes towards Moroccan immigrants than Canadians do towards Asian immigrants–who are in fact seen as the 'model minority' (Little, 2016)–it will be interesting to investigate whether these contrastive attitudes towards these two ethnic groups (and, by extension, their speech) have different effects on speech processing and evaluation. In other words, do negative attitudes towards a specific group exacerbate the potential effects of an expectation mismatch effect on reverse linguistic stereotyping?

5.3. Methods Experiment 3

5.3.1. Stimuli

This experiment was divided into three tasks: (1) a speeded sentence verification task to determine the effects of accent and visual speaker ethnicity on response latencies, (2) an accentedness rating task, and (3) a credibility assessment task to test whether people were less likely to believe a fact that had been spoken by an accented speaker. For the speeded sentence verification task, sixty-four statements (32 true, 32 false) were recorded. Most of the stimulus sentences had been directly translated into Dutch from the previous experiment. Because of syntactic differences between English and Dutch, some of the sentences had to be replaced to ensure that participants needed to hear the entire sentence before they could judge its veracity. For instance, the sentence *'you can keep dogs as pets'* would translate to '*je kan honden als huisdier houden' [you can dogs as pets keep]* in Dutch. The sentence-final verb in this example has very high predictability, which increases the risk of participants responding before the entire sentence has been spoken, so sentences like these were replaced with entirely different sentences. The revised list of sentences was vetted informally for potential ambiguities by two native Dutch speakers. These sentences ranged between three and eight words, with a mean length of 5.4 words per sentence. They are listed in Appendix I.

The video recording was done by the researcher on location on an iPad that had been suspended at face-height. Since the recordings were made in a location that was convenient for the speakers, some recordings contained occasional background noises, despite every effort to make the environment as quiet as possible. The speakers were asked to say aloud the 64 true/false sentences and 43 trivia sentences while looking in the camera. The researcher indicated when a sentence had to be repeated because of hesitations or errors. While the original plan had been for the participant to read the sentence, look up into the camera, and briefly pause before recounting the sentence from memory, this proved to be challenging for some of the MFD participants. Therefore, the decision was made to have all speakers repeat after the researcher, as this still ensured that the participants did not appear to be reading. This may have caused the speakers to converge towards the accent of the researcher, but this effect ideally would have been the same across all speakers. After recording the entire dataset, fourteen sentences were

removed due to intelligibility issues and unexpected background noise, resulting in a total of 50 sentences (25 true, 25 false).

Background noise in the audio recordings was attenuated in *Audacity* (Audacity Development Team, 2018) using the 'noise reduction' function. The audio files were then spliced into sentence-length files and normalized for peak amplitude. Word-final fricatives were cut short at a zero crossing, and any noticeable hesitations between words were shortened. For each sentence, the median of the four speakers' renditions of the same sentence was calculated, after which the durations of the four files were digitally compressed or lengthened in *Praat* (Boersma & Weenink, 2018) so that all versions of the same sentence had the same length, thereby reducing the potential confounding effect of durational differences on response times. A 500 ms silence was then added to the beginning of each audio file.

As in the previous experiment, there were noticeable differences in amplitude between the four speakers despite normalization, so the audio files' LUFs[26] were adjusted with the *Audacity* plugin ReplayGain to 89 dB. Then, the same multi-talker-babble file that had been used in the previous experiment was mixed with the audio files, with the same signal-to-noise ratio of +5 dB. Next, the video files were cut into sentence-length clips, stripped of their audio track, and synced with the four speakers' audio files in *Adobe Premiere Pro* (Adobe Systems, 2019). This video editing program allows speed manipulation of videos on a millisecond-specific timeline. By speeding up or slowing down sections of the video, the audio tracks were synced with the corresponding videos. Finally, a blur (specifically, a feathered mosaic) was manually placed over the mouth and jaw area of each speaker to cover up any remaining syncing dysfluencies. Mask tracking was enabled, meaning that the blur size and location were automatically adjusted throughout the video to keep the lips and jaw movements concealed at all times, regardless of head movements. This was done for a total of 800 videos. For each of the 50 sentences, the four different audio versions were paired with the four video versions of that same sentence, resulting in 16 unique versions (4 speakers x 4 voices) of each sentence.

[26] LUFS = Loudness Units relative to Full Scale

To ensure a balanced design, sixty-four unique experiment configurations were created, where each participant saw only four of the possible sixteen guises. Once again, the test items were blocked by speaker, such that listeners heard twelve or thirteen sentences spoken by the same guise before hearing the next speaker. The sentences themselves were counter-balanced across the guises, just as speaker order was counterbalanced. Sentence order within blocks was randomized. An additional sixteen unique configurations were created for the Audio-Only condition. As there were eighty-four listener participants, some configurations were used twice. Participants were assigned to one of three categories: all congruent, all incongruent, half congruent-half incongruent, or Audio-Only.

While the design of the response time task remained the same as in Experiment 2, this experiment's accentedness rating task was modified; listeners were now reminded of how the various guises had sounded before being asked to assign an accentedness rating; they were first shown a still of each guise, along with a recording of the voice they had heard with that guise saying three sentences. This ensured that the ratings were less dependent on the listeners' recall abilities than had been the case in the previous experiment. The rating scale was once again a seven-point Likert scale (1 = native/Dutch-sounding, 7 = foreign/not Dutch-sounding).

Credibility assessment task

Finally, the added credibility assessment task required participants to indicate the believability of a series of trivia statements. The objective of this task was not to test the participants' worldly knowledge, but instead to investigate whether MFD speakers were perceived as less credible than the native Dutch speakers. This task was motivated by Lev-Ari and Keysar's (2010) finding that native speakers of English judged trivia statements as more true when uttered by a fellow native speaker of English as opposed to a foreign-accented one. Note that the actual observed difference in credibility between native speakers and heavily-accented speakers was 0.75 cm on a 14 cm scale (i.e., 5.4%), while the difference between native speakers and mildly accented speakers was only 0.64 cm (4.6%). The authors drew a causal link between processing ease and credibility, concluding that since accented speech reduces processing fluency, it is consequently perceived as less believable. This conclusion garnered a surprising amount of media attention in the Netherlands, with large newspapers running headlines such as 'Your

Dunglish could mess up a deal' (Kist, 2015a), 'Steenkolenengels[27]: not so very believable,' ("Steenkolenengels," 2015) and 'Unnatural accent? One-nil down' (Kist, 2015b). This reaction was fascinating, not in the least because none of the foreign-sounding speakers featured in the study had in fact been Dutch. One of the articles misrepresented the study's findings altogether as evidence that "bad Dunglish," such as literal translations of Dutch expressions into English (e.g., 'go your gang' as a literal translation of the Dutch equivalent of 'go ahead,') could cost people business deals. The fact that the study focused on mild and heavily accented (but grammatically correct) sentences was lost on many journalists as well.

Several studies have failed to replicate Lev-Ari and Keysar's (2010) findings. So far, the credibility of foreign accents has been unsuccessfully tested on native English listeners (Souza & Markman, 2013), Swiss-German and French listeners (Stocker, 2017), and Italian listeners (De Meo, 2012; De Meo, Vitale, Pettorino, & Martin, 2011).[28] Even regional and local accents were found not to differ in credibility (Frances, Costa, & Baus, 2018). The only study that appears to have uncovered a similar trend was conducted by Hanzlíková and Skarnitzl (2017), on non-native listeners. When comparing the credibility ratings assigned to 'native' speech (provided by British and American speakers) with the ratings given to 'non-native' speech (provided by Czech speakers and a mix of other non-native speakers), they discovered a significant difference between the two groups. However, it is worth pointing out that the difference was a mere 0.4 points on a 7-point Likert scale (i.e., 5.7%), and is thus far from a strong endorsement of Lev-Ari and Keysar's (2010) findings. In addition, when the various accents were compared separately (rather than collapsing them into a 'native' and 'non-native' category), this reported difference in believability between the native and non-native voices became less straightforward, as the difference between the non-native Czech and native American-accented voices was now no longer significant. While the authors concluded that their results broadly confirmed Lev-Ari and Keysar's (2010) findings, the small difference in credibility between the two speaker groups (0.4 points) and the inconsistent results when accents were looked at separately is cause for some skepticism.

[27] Translation: 'Dunglish'

[28] See Hanzlíková and Skarnitzl (2017) for a more detailed overview and description of these three studies.

Regardless, since the topic of accent credibility was of considerable interest to the general public in the Netherlands, an accent credibility assessment task was added to the present experiment to test whether Dutch listeners would follow the same trend as observed by Lev-Ari and Keysar (2010) and Hanzlíková and Skarnitzl (2017). Instead of using a recognizably 'foreign accent' such as English-accented Dutch, however, this will be the first replication experiment to compare the purported credibility of an ethnolect (MFD) against the credibility of a Standard Dutch accent.[29] The results were further extended by adding visual speaker information to test whether specific guises would elicit different credibility ratings.

The speakers also recorded 54 trivia sentences for the accent credibility assessment task. Some had been directly translated from Lev-Ari and Keysar (2010), who graciously agreed to share their stimulus sentences. Others had been collected from trivia websites. Not all sentences were included in the experiment, as some (1) proved to be too long, (2) had intelligibility issues, or (3) simply caused difficulty for the speakers. For instance, the sentence "paars licht wordt niet efficiënt door planten geabsorbeerd" [purple light is not efficiently absorbed by plants] proved to be a tongue twister for native Dutch speakers and MFD speakers alike. In the end, 43 trivia sentences were included (21 true, 22 false). They are listed in Appendix J.

Because creating the stimulus videos was extremely labor intensive, the decision was made not to pair the trivia sentences with videos. Instead, they were shown in conjunction with a photograph of the speaker. The trivia sentences were normalized for peak amplitude across speakers, but the endings were not clipped, as reaction times were not being collected. Background noise was attenuated in the audio files with the 'noise reduction' function in *Audacity* (2018).

5.3.2. Stimulus Design

A total of fifteen women lent their voices to this project. One participant provided a high-quality voice recording by iPhone, while the other fourteen women were video recorded for this experiment. To keep regional provenance constant across speakers, all

[29] Specifically the 'Randstad accent,' which is the variety with the greatest prestige (Grondelaers et al., 2015, p. 197)

speakers were recruited in the Randstad; a "heavily urbanized area in the west (comprising the major cities Amsterdam, Rotterdam, The Hague, and Utrecht) which is the political and socio-economic hub of the Netherlands" (Grondelaers et al., 2015, p. 195). The area is typically not associated with accentedness. To select four voices and faces for the experiment, a short online survey was created on *SurveyMonkey.* The participation link was posted on Facebook and sent around by email. Two versions were available; one for Dutch speakers (*n* = 18), and one for non-Dutch speakers (*n* = 42). In both versions, participants heard twelve of the speakers say a few of the Dutch sentences, after which they had to estimate their respective ages. They also provided age estimates for screenshots of the speakers' videos (without knowing which voices belonged to which faces). Participants in the Dutch version were additionally asked to guess where they thought the speakers had grown up, e.g., by listing a specific province in the Netherlands, or a country.

Because the questions about the speakers' respective ages and provenance were open-ended, there was considerable variety in people's responses, suggesting that this information was difficult to discern[30]. For instance, one of the MFD voices had age estimates ranging from 19 to 58. Indeed, many participants indicated that they had found the task extremely challenging, as they had had no idea for most of the speakers. This feedback, combined with the results, allowed me to choose the four stimulus voices that I thought best represented MFD and SD, without having perceived age differences between the four speakers potentially confounding the results. The two MFD voices were chosen to reflect two different degrees of accentedness. Their mean estimated ages were 35 and 36, while the mean estimated age for the selected Dutch voice was 30.

Contrary to the lack of agreement in perceived age based on the speakers' voices alone, there was more consensus when the speakers' faces were shown instead. Apart from one speaker, all Dutch speakers were noticeably older than most of the Moroccan speakers (see Appendix K for images of the twelve speakers). This posed a problem, as the Moroccan and Dutch women had to be roughly similar in both age and appearance to ensure that they were believable guises. Campbell and McCullough (2015) discussed this exact methodological challenge, explicitly calling on researchers to ensure that their

[30] 31% of the listeners identified the Moroccan speakers as Moroccan, 22% as coming from Amsterdam, 13% as Turkish, and 34% had chosen miscellaneous 'foreign' labels.

manipulated voice and face matches were believable. Therefore, to get more age-appropriate white visual guises that all looked like they were in their thirties, two additional Dutch women were video recorded in Vancouver.

To avoid having 'perfect fit' pairings where the voice and face truly belonged together, the original voices from the two women who had been videotaped in Vancouver were excluded from the experiment. This meant that the experiment was still short of one 30-year-old-sounding Dutch native voice. This voice came from a 29-year-old woman from the Hague in the Netherlands who volunteered to send a recording of her voice by email.

5.3.3. Speakers

The four voices selected for this study belonged to two female native Dutch speakers, and two female speakers of MFD. One of the native Dutch speakers (dubbed 'Native1') grew up in a small village relatively close to Rotterdam (age 30), while the speaker dubbed 'Native2' grew up in Groningen (age 29) but moved to the Randstad (specifically the Hague) at the beginning of her studies. Despite their different regional backgrounds, neither had strong regional markers in their accents that set them apart from people in the Randstad, which is the geographical area where most of the data collection took place. The two Moroccan speakers selected for this experiment had both been born in Morocco but had moved to the Randstad area in the Netherlands at a young age. The speaker whose voice was later dubbed 'Foreign1' was 21 years of age, while the speaker whose voice was dubbed 'Foreign2' was 51 years old. Assessment by the researcher indicated that Foreign1 had a much heavier accent than Foreign2. The intonational patterns of Foreign1 also sounded a bit unnatural at times. This was unavoidable because the researcher had to cut the sentence into smaller sections for the speaker to be able to repeat after her.

The following four faces were selected to create the sixteen guises:

Moroccan1 Moroccan2 White1 White2

Figure 5.1 The faces used in this experiment

5.3.4. Listeners

A total of eighty-six listeners participated in the experiment. As in the previous two experiments, no restrictions were placed on their ethnic backgrounds. All were native speakers of Dutch, who had reported normal hearing. The data from two participants was removed from the analysis because they reported a hearing problem upon completion of the experiment. Of the remaining 84 participants, 64 completed the Audio-Visual experiment (48 female, 16 male). Twenty-two of the participants (25.6% of the total population) completed the Audio-Only experiment (13 female, 9 male) as the control group.[31] All but five participants were Caucasian.

Twenty-seven of the participants were recruited in Nijmegen–a smaller city in the far east of the Netherlands–through the Radboud University Research Participation System. The other participants were recruited in and around Amsterdam through the social and professional networks of my friends and family members. To the best of my knowledge, none had a background in linguistics. Because of this snowball approach, a diverse sample of participants was reached with varying levels of education and income–something which would have been far more difficult if only university students had been recruited. Participants represented a variety of social backgrounds, and ranged in age between 17 and 68, with a mean age of 37. All participants were offered €5 for their time, but many declined remuneration. See Figure 5.2 below for the distribution of ages from the listener sample (84 participants).

[31] This percentage of participants assigned to the control group (i.e., Audio-Only) falls within the range that White (2018) considers a good compromise between getting the most number of participants exposed to the treatments while keeping the loss of statistical power as small as possible.

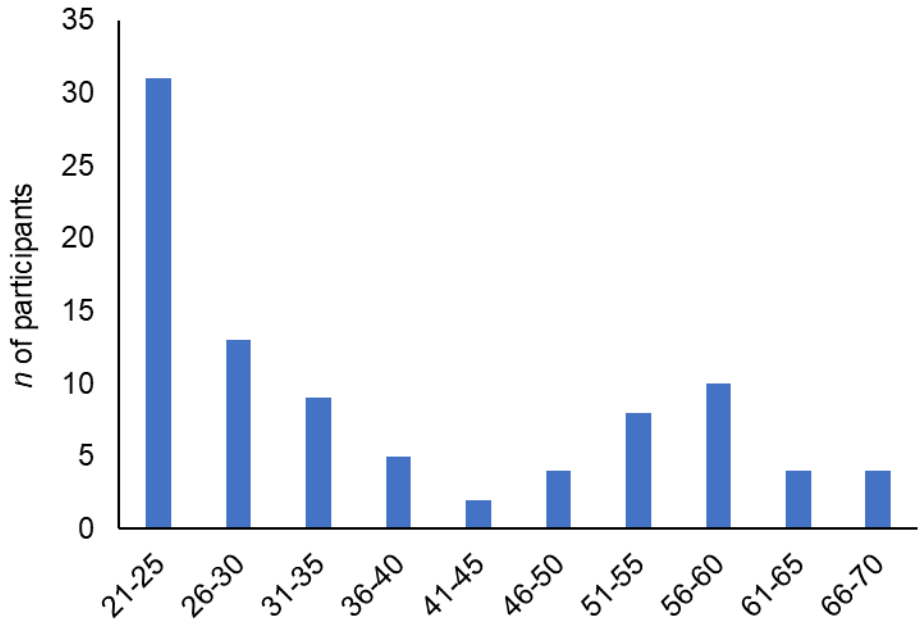


Figure 5.2 Age distribution of participants

5.3.5. Procedure

Since the stimulus presentation program used in the previous experiment had raised issues with retrieving videos over a Wi-Fi connection, the decision was made to use a different stimulus presentation program for this experiment. Replicating the design with different software would also help to clarify whether the observed increase in RTs to Audio-Visual sentences–as reported in Chapter 4–was truly an artefact of a software malfunction, or reflective of an actual trend. Experiment 3 was run on *Cedrus Superlab 5* (2017 Windows edition), a millisecond accurate stimulus presentation program that makes use of a physical response pad to measure reaction times. The RB-740 response pad was connected to the USB port of an ASUS laptop (Windows 10). A pair of high-quality noise-cancelling headphones (Model: Sony WH1000XM2/B XB) was also connected to the laptop and calibrated to the noise profile in the room to optimize ambient sound cancellation. The volume was pre-set to a comfortable listening volume.

All participants were tested individually in a silent room at a location convenient for them. The researcher was present throughout the entire experiment. Participants were informed that their first task was a response time study, and that speed and accuracy were equally important. They were explicitly instructed to keep their index finger close to the response pad. For the AV condition, they were also told to keep their eyes focused on the

speakers' faces. After signing the informed consent form, they were seated behind the laptop and response pad.

For task familiarization, the participants completed six practice items which featured the experimenter herself (Caucasian female speaking with a Standard Dutch accent). To help convince the listeners that the face–voice combinations they saw were real, the videos were not manipulated, i.e., her own voice and face were used. The videos were presented against a black background, with 0.5 s between participant responses and the presentation of the next stimulus. For the first task, three buttons were activated on the response pad (See Figure 5.3 below). The button with the green cover was to be pressed when the sentence was true, while the button with the red cover was for false sentences. The button that featured a question mark was to be selected when participants did not understand what had been said. This was a forced choice experiment, i.e., participants had to hit any of the three activated buttons before they could move on. Any accidental pressing of the inactivated buttons (1, 2, 4, 6) was recorded, but did not result in the presentation of the next stimulus. Any button presses before the video or audio had finished playing would be recorded as well, but the video/audio would play out before the next audio file was presented. After the video had ended, the last frame of the video remained on the screen until a button was selected. In the Audio-Only condition, a fixation cross was shown instead of a video. The rating of all 50 sentences took an average of six minutes.

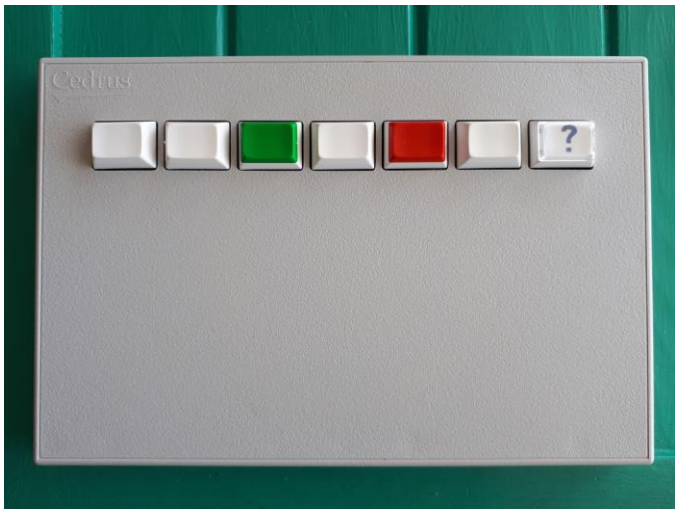


Figure 5.3 The RB-740 response pad configuration for task 1

For the second task, all seven buttons were activated. Participants heard the four guises say ten or eleven trivia statements each, which added up to 43 statements total. Trivia statements included phrases such as ‘ants don’t have lungs,’ or ‘it rains diamonds on Jupiter and Saturn.’ See Appendix J for the full list of trivia sentences and their translations. Participants were told that speed was no longer the primary goal, but that they nonetheless should not think too long about each statement, as the task was meant to capture their “first impression of whether they thought the trivia statement was true or false.” The word ‘credibility’ was purposefully avoided in the instruction, so as not to give the goal of the task away. They were also told that in addition to indicating whether they thought the statement was true (left side/buttons 1-3) or not (right side/buttons 5-7), the number they picked within that range indicated the level of confidence in their own judgment. See Table 5.1 to see what each button represented.

Table 5.1 The seven options in task 2 to indicate trivia sentence veracity & rater confidence

	Dutch	English translation
1	Zeker waar	Definitely true
2	Waar	True
3	Waarschijnlijk waar	Probably true
4*		
5	Waarschijnlijk onwaar	Probably false
6	Onwaar	False
7	Zeker onwaar	Definitely false

* Didn’t hear / Neutral

Each button was numbered, and six color-coded descriptors were added around the buttons on the response pad (see Figure 5.4). The same color-coded 7-point scale was also shown on the screen after each sentence finished. Participants were verbally instructed to use button 4 if they had not understood what had been said, or if they were truly neutral–although they were strongly encouraged to pick either side of the scale.

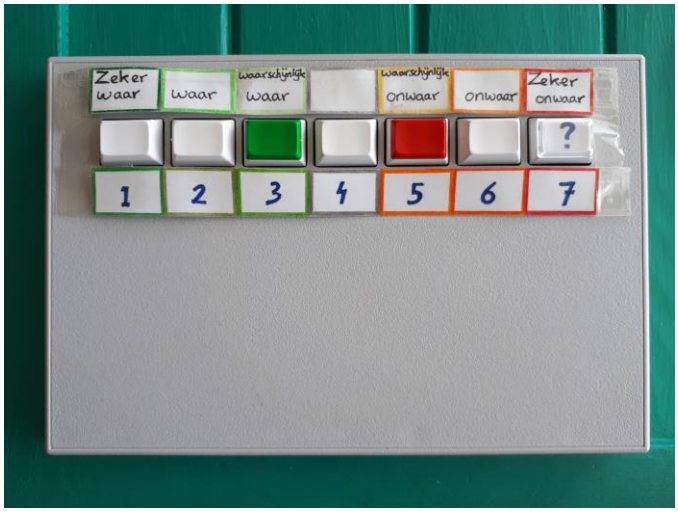


Figure 5.4 The RB-740 response pad configuration for task 2

The third and final task measured perceived accentedness for each speaker. The participants were played three randomly chosen recordings for each speaker as a reminder of what they sounded like before they assigned an accentedness rating on a Likert scale from 1 ('Dutch-sounding/native speaker') to 7 ('foreign-sounding/non-native speaker'). This Likert scale was shown on the screen after each recording. Participants in the Audio-Visual condition also saw a still of the speaker they had seen with that voice.

Similar to the software presentation system that had been used in the previous study, *Superlab 5* also measured response times from the moment the stimulus audio/video started playing. Since the four speakers' renditions of the same sentence occasionally varied slightly in duration, the stimulus files' respective durations were subtracted from the reaction times to calculate the actual reaction times. In addition to response times, *Superlab 5* also recorded the listeners' responses (true/false/???) to each sentence. It then compared this to the correct response and assigned a value of 0 to a '???' or a wrong response. A correct response received a value of 1. In the accent credibility assessment task, the participants' reaction times were also recorded (even though they had been told that speed was not important), together with the credibility score they had assigned to each trivia sentence.

Following Campbell-Kibler and McCullough's (2015) call to check the believabilty of the face and voice combinations, the experiment concluded with two questions. First, participants were asked what they thought the experiment was about. Many mentioned accents in relation to processing speed, but none of the participants was able to correctly

identify the purpose of the study. The follow-up question focused on which speaker they had found most difficult. Leading questions such as 'did you notice anything strange about the speakers?' were avoided. From the participants' answers it became clear that virtually none had realized that the guises' voices had been manipulated. Only two participants indicated that they thought the voices did not belong to the speakers. The others truly believed that the voices had belonged to the women, with many commenting that the 'Russian lady' or 'East-European lady' (i.e., the incongruent White face–Moroccan accent pairing) was the most difficult to understand. This was a recurring observation; the Moroccan accent was easily identified when it had been paired with a congruent ethnically Moroccan speaker, but was misidentified as East-European when paired with a White face.

5.4. Predictions

Considering the fact that Lev-Ari and Keysar's (2010) findings have not been replicated with the exception of Hanzlíková and Skarnitzl (2017), whose findings were arguably not very robust, it is unlikely that listeners are more inclined to believe trivia statements said by a SD speaker compared to a MFD speaker. If there is a direct link between increased processing difficulty with accented speech and perceived credibility, then the MFD speakers should be perceived as less credible, as their speech is predicted to be more difficult to process. Because there have been so few studies testing this purported credibility effect, it will be helpful to see whether Lev-Ari and Keysar's (2010) findings can be replicated.

Based on the observations made in Chapters 3 and 4 regarding the increased processing cost of foreign-accented speech, it is expected that the Moroccan-accented voices will take longer to process than the native Dutch voices–irrespective of their purported ethnicity. Although the previous study in Chapter 4 yielded no effect of visual speaker ethnicity on comprehensibility, this study could potentially produce a different result because of its different cultural setting, target language, and the listeners' overall more negative attitudes towards the immigrant group under studyresponse. Yet as there have been conflicting past results, no prediction can be made about the effect of visual speaker ethnicity on accentedness ratings and processing speeds at this time.

Based on Munro and Derwing's (1995) findings regarding the increased processing cost associated with foreign-accented speech, we would expect the Moroccan voices to elicit longer response times than the Standard Dutch voices. If there is an effect of visual speaker ethnicity on comprehensibility, then response times could potentially be shorter in a congruent as opposed to an incongruent condition. Yet if the data shows longer response times to the Moroccan guises compared to the White guises (irrespective of voice), then the data would be more in line with the reverse linguistic stereotyping hypothesis.

If visual speaker ethnicity is found to influence accentedness ratings as well, two specific types of outcomes would provide evidence for the exemplar theory and RLS, respectively: if incongruent pairings were to receive 'worse' (i.e., higher) accentedness ratings than the congruent pairings, then this would support the exemplar theory. Yet if the Moroccan guises would receive higher accentedness ratings than the White guises, regardless of their accent, then this would be more in line with the predictions of the reverse linguistic stereotyping hypothesis.

Within the context of this experiment, the incongruent face–voice combinations are assumed to be White guises speaking MFD, or Moroccan guises speaking SD. While MFD is also spoken by people outside of the Moroccan-Dutch community, its reputation as a vernacular and the 'language of the street' make it less readily associated with White speakers, let alone female ones. Similarly, ethnically Moroccan women speaking SD is not a very common occurrence either, as most second and third-generation Moroccan immigrants actively choose to maintain their MFD as an identity marker (Nortier & Dorleijn, 2017)

Finally, based on the observation that older listeners have more difficulty understanding foreign-accented speech than younger listeners–possibly due to differences in hearing acuity and working memory capacity (Ingvalson et al., 2017)–the MFD speech is also expected to entail longer RTs for the older participants than for the younger participants, even though MFD is an ethnolect rather than a foreign accent. Since there is also evidence suggesting that older adults are more prejudiced than young adults (Gonsalkorale et al., 2009; Stewart et al., 2009; Von Hippel et al., 2000), it is furthermore hypothesized that older participants will (1) take longer to respond to, and (2) assign higher accentedness ratings to the Moroccan guises than to the White guises.

5.5. Results

5.5.1. Analysis I: Response times

The first analysis investigated the effect of visual speaker ethnicity on listeners' reaction times to MFD and SD. In this analysis, the four speakers' voices were each treated as a level of the variable *Voice ID.* 'Native1' and 'Native2' represent the two native Dutch speakers, while 'Foreign1' and 'Foreign2' describe the two MFD voices. Since the four speakers should not be assumed to be representatives of their respective two accents, this analysis analyzed each of the four voices separately, rather than collapsing them into a 'Moroccan' and 'Dutch' accent category[32]. To keep the number of levels manageable, and because any differences between speakers of the same ethnicity were not the main focus of this analysis in the analysis, the four speakers were categorized into either 'White' or 'Moroccan.' An additional 'NoFace' level was added for respondents in the Audio-Only condition. Just as in the previous experiment, the response times to incorrectly identified sentences were removed from the analysis. Since many more MFD sentences were incorrectly identified (n = 602) than SD utterances (n = 271), a disproportionate number of response times to the MFD sentences had to be discarded. This may have caused the model to under-estimate the differences in RTs between the speakers, as many of the removed MFD utterances had longer response times than the SD utterances. See Table 5.2 below for an overview of how many response time data points were removed for each face–voice combination (total *n* = 873).

Table 5.2 Counts of incorrect identifications per level of *Voice* and *Face*

	Voice			
Face	Native1	Native2	Foreign1	Foreign2
White1	38	20	87	38
White2	22	19	73	37
Moroccan1	20	22	69	30
Moroccan2	31	22	80	27
Audio-Only	41	36	125	36
Sum of discarded sentences per voice	**152**	**119**	**434**	**168**

[32] A separate analysis was run to check whether the results changed in a model that *did* collapse the voices into a 'native/standard' and 'non-native/ethnic' accent, as has been done in other studies. Since the findings were the same, they will not be discussed in this chapter.

A fitted vs. residuals plot and a Q-Q plot showed that the residuals in the initial model violated both the constant variance and normality assumptions of regression, so the response times were log-transformed to improve the model fit. Eight observations with a marginal studentized residual of > 4 were identified as outliers, but because their removal did not change the findings, the analysis of the entire dataset will be reported here. Since the previous analysis had found a relationship between Log(RT) and *Trial*, this was examined first. However, contrary to what had been observed in the previous analysis described in section 4.3.1, there was no clear upward or downward pattern in Log(RT) in relation to *Trial*, nor was there evidence of a difference between the Audio-Visual and Audio-Only conditions. Both these findings lend credence to the hypothesis that the previously reported increase in response times as the experiment progressed may indeed have been the result of a software malfunction. Figure 5.5 below shows the spread of RTs without the outliers. Only the RTs to correct responses were plotted.

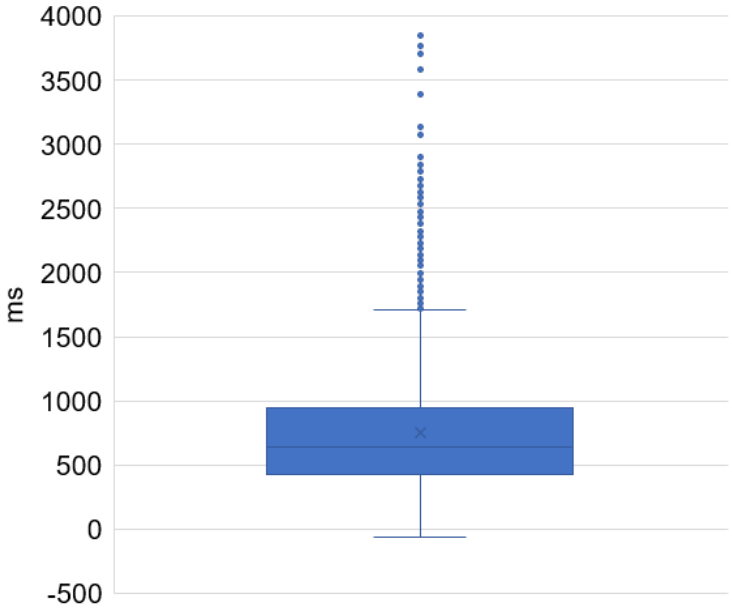


Figure 5.5 Boxplot showing the spread of raw response times in ms

The linear mixed effects model fit to the response time data included several terms. The first fixed effects variable *Face* had three levels: White, Moroccan, and Audio-Only, as I was not interested in any differences in response times between two speakers of the same ethnicity. *Voice ID* had four levels: Native1, Native 2, Foreign1, and Foreign2. The term *Veracity* indicated whether the sentence was true or false, while the term *Age*

represented a participant's age, and was treated as a continuous variable. The model did not include a *Trial* effect, as no terms involving it were significant.

The fitted model also contained the following, normally distributed, random effects with random intercepts: *Participant* helped account for any participant-specific main effects on response times. Treating it as a random effect facilitates generalizing the results beyond this study. *Sentence* served as a random intercept identifying which sentence a participant was presented with. This helps account for sentence-specific effects on response time and allows the results of the study to be generalized to sentences beyond the sixty-four that were used in this study. *SV* was a random intercept identifying which *Sentence–Voice ID* combination was shown to a participant. Since there are multiple observations for each *Sentence–Voice ID* combination, it is possible to estimate if there are any interaction effects in the data. *SF* was a random intercept identifying which *Sentence–Face* combination a participant was shown. Since there are multiple observations for each *Sentence–Face*, it is possible to estimate if there are any interaction effects in the data. *SFV* was a random intercept identifying which *Sentence–Face–Voice* combination a participant was shown. Again, since there are multiple observations for each *Sentence–Face–Voice*, it is possible to estimate if there are any interaction effects in the data. As in the previous two experiments, the lmerTest package (Bates, Mächler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2017) was used to code dummy variables to produce beta estimates, while the emmeans package (Lenth, 2019) was used to generate the tables that contain treatment group means (by using the results and coefficients of the lmer model fit). Appendix L shows a summary of the linear mixed effects model's estimated regression coefficients and the variable coding scheme that was used.

An ANOVA was carried out on the fitted model in *R* using the Kenward-Rogers degrees of freedom approximation method with Type 3 Sums of Squares.

Table 5.3 Type III Analysis of Variance Table with Kenward-Roger's method

Fixed effect	Sum of squares	Mean squares	NumDF	DenDF	*F*	*p*
Face	0.656	0.328	2	110	1.572	0.212
Voice ID	13.278	4.426	3	946	21.327	< .0001
Veracity	0.862	0.862	1	48	4.154	0.047
Age	1.215	1.215	1	85	5.854	0.018
Face * Voice ID	1.333	0.222	6	397	1.071	0.380
Face * Veracity	0.317	0.158	2	97	0.763	0.469
Voice ID * Veracity	0.165	0.055	3	141	0.264	0.851
Voice ID * Age	2.571	0.857	3	3217	4.130	0.006
Face * Voice ID * Veracity	0.754	0.126	6	265	0.605	0.726

As can be seen in Table 5.3 above, there were no significant *Face*, or *Face * Voice ID* effects, thus providing evidence against the existence of ethnicity-based bias or incongruency effects on response times. There was however a significant *Voice* effect, as well as a significant *Veracity* effect, suggesting that there were differences in RTs between the four voices, and that the veracity of the sentence also played a role. No higher order interaction terms involving *Face, Voice ID* or *Veracity* were significant. There was also a significant *Age* term, as well as a significant *Voice ID * Age* term, providing evidence of a participant age effect in the data, and that this effect varies depending on the voice presented.

Next, we will examine the model estimated medians (instead of the means because of the log-transformation) and further examine what effects were present in the data. Although the omnibus test did not yield a main effect for *Face,* it was still of interest to this study to calculate the model-estimated medians for each level of *Face*. This was done with the emmeans package in *R* (Lenth, 2019). As can be seen in Table 5.4 below, despite the audio-only group having a noticeably lower median than the participants in the AV condition, the analysis did not find these differences to be significant.

Table 5.4 Median response times in ms at each level of *Face*

				95% Confidence Interval	
Face	Median RT	SE	df	Lower	Upper
Audio-only	598	43.919	113	516.802	691.438
Moroccan	688	35.116	138	621.671	760.781
Dutch	688	35.160	138	621.896	761.177

* Conditions that share the same number in the 'Group' column are not significantly different from one another

Next, the main effect of *Voice ID* was examined in more detail by testing whether there were any significant differences between the four individual voices. Since there was a significant *Voice ID * Age* interaction term, Tukey-Kramer post-hoc tests were conducted at different values of *Age*. The decision was made to examine the voice differences at the 1st (25%), 2nd (50%) and 3rd (75%) quartile values that *Age* took on in the data in order to get estimated measures for 'young' (age = 23), 'middle-aged' (age = 31) and 'old' (age = 53) listeners. Since the *Voice* effect did not appear to change with age (i.e., the groupings were identical for all three quartiles), only the table for the 1st quartile will be shown below (Age = 23); the results of the comparisons for the other two *Age* quartiles can be found in Appendix N. As can be seen below, the two foreign-accented voices had significantly longer RTs than the two native voices, with Foreign1 taking significantly longer to process than Foreign2.

Table 5.5 Median response times in ms at each level of *Voice ID* for *Age* = 23

				95% Confidence Interval		
Voice ID	Median RT	SE	df	Lower	Upper	Group*
Native1	504	29.8	180	449	567	1
Native2	554	32.7	178	493	623	1
Foreign2	666	39.5	181	592	748	2
Foreign1	763	46.5	200	677	861	3

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Just as in Chapter 4, a *t*-test was used to determine whether median response times in the Audio-Visual condition were longer than in the Audio-Only condition. However, contrary to what had been found in Chapter 4, this contrast did not yield significant evidence (p = 0.079) that the ratio of median response times for AV and Audio-Only sentences was different from 1, allowing us to conclude that there were no significant differences in RTs between the AV and AO conditions.

Table 5.6 Contrast of median RTs for AV and AO conditions

			95% Confidence Interval			
Contrast Ratio	SE	df	Lower	Upper	*t* ratio	*p*
1.150	0.091	86	0.984	1.350	1.778	0.079

Between-subjects analysis

Each participant was shown four guises, which gave them a frame of reference, and enabled data collection on both within- and between- subject variation. While this experiment yielded no effect of guise ethnicity, some previous studies that exposed participants to a single guise did (e.g., Gnevsheva, 2018; Kang & Rubin, 2014; McGowan, 2015; Rubin, 1992). To test whether this different result may have been due to experiment design, the participants' reaction times to only their first guise were separately analyzed as if they had only been exposed to just that one guise. The same model that was described above was fit to this sub-set of the data. Below are the results:

Table 5.7 Type III Analysis of Variance Table with Kenward-Roger's method

Fixed effect	Sum of squares	Mean squares	NumDF	DenDF	*F*	*p*
Face	0.205	0.102	2	69	0.584	0.561
Voice ID	0.993	0.331	3	64	1.885	0.141
Veracity	0.299	0.299	1	240	1.701	0.193
Age	0.103	0.103	1	69	0.587	0.446
Face * Voice ID	0.396	0.066	6	64	0.376	0.892
Face * Veracity	1.154	0.577	2	354	3.282	0.039
Voice ID * Veracity	0.572	0.191	3	597	1.085	0.355
Voice ID * Age	0.444	0.148	3	64	0.843	0.475
Face * Age	0.016	0.008	2	67	0.046	0.955
Veracity * Age	0.067	0.067	1	716	0.384	0.536
Face * Voice ID * Veracity	4.059	0.676	6	629	3.851	0.001
Face * Voice ID * Age	0.528	0.088	6	63	0.502	0.805
Voice ID * Veracity * Age	0.349	0.116	3	714	0.663	0.575
Face * Veracity * Age	0.538	0.269	2	681	1.532	0.217
Face * Voice ID * Veracity * Age	2.691	0.449	6	693	2.555	0.019

The significant *Age * Voice ID * Face * Veracity* term in this model indicates an age effect in the data, and that this effect varies depending on which *Face–Voice-Veracity* combination a participant was shown. Because of this, the subsequent incongruency tests were carried out at the 1st, 2nd, and 3rd quartile values of *Age*. However, neither the overall incongruency effect hypothesis test, nor the voice-specific incongruency contrasts indicated statistically significant evidence of any incongruency effects in the data. Estimated medians for each *Face–Voice-Veracity* combination were also calculated at the three quartile values, but since so many simultaneous comparisons were made, the *p*-values had become too conservative to provide a reliable insight into potential interactions.

In short, this analysis of the participants' response times to a single guise (as would have been the case in a between-subjects study design) rendered no different result, as visual speaker ethnicity was not found to affect response times.

Discussion Analysis I

The primary purpose of this analysis was to investigate whether certain combinations of *Face* (Moroccan, White) and *Voice ID* (Foreign1, Foreign2, Native1, Native2) entailed significantly different response times. If the unexpected face–voice pairings had taken the longest to process, this would have provided evidence for the exemplar theory. If the Moroccan guises had elicited the longest response times instead (irrespective of their accent), this would have provided support for the reverse linguistic stereotyping hypothesis. Interestingly, the data from this study did not provide support for either theory, as the participants were not found to respond differently to any of the manipulated guises. Similarly, a separate analysis of the participants' responses to only the very first guise provided no evidence of listener bias or incongruency either, as there were no significant differences between the congruent and incongruent guises, nor did guise ethnicity affect response times.

The prediction that participants would take significantly longer to respond to the MFD voices compared to the SD ones (irrespective of the face they had been paired with) was borne out in the data. This finding is similar to what was observed in Chapter 4, where the Japanese-accented voices also entailed longer RTs than the native voices. Where there were no differences between the two native Dutch voices, the two Moroccan-flavored voices did entail significantly different RTs.

Finally, contrary to what was reported in Chapter 4, listeners in the Audio-Visual condition had equally fast RTs as those in the Audio-Only condition. There was also no effect of *Trial,* as participants did not become faster or slower as the experiment progressed. These two findings strongly suggest that the previously-reported effect of *Trial* and the differences between the Audio-Visual and Audio-Only conditions in Chapter 4 may have been caused by a malfunction of the stimulus presentation software, rather than an actual effect.

5.5.2. Analysis II: Accentedness ratings

Next, the accentedness scores were analyzed. A linear mixed effects model was fit to the data, with *Face* and *Voice* as fixed effects, *Participant* as a random intercept, and *Age* and various interactions that can be seen below in Table 5.8 as fixed effect covariates. The ANOVA that was carried out on the fitted model revealed a significant effect of *Voice ID*, and a significant *Face * Age* interaction, in addition to a three-way interaction between *Face, Voice ID*, and *Age*.

Table 5.8 Type III Analysis of Variance Table with Kenward-Roger's method

	Sum of squares	Mean squares	Num DF	Den DF	*F*	*p*
Face	3.154	1.577	2	161	2.205	0.114
Voice ID	329.986	109.995	3	246	154.476	< .0001
Age	1.698	1.698	1	83	2.385	0.126
Face * Voice ID	4.258	0.710	6	298	0.996	0.428
Face * Age	5.890	2.945	2	161	4.119	0.018
Voice * Age	0.560	0.187	3	246	0.262	0.853
Face * Voice ID * Age	9.540	1.590	6	297	2.232	0.040

Because of the significant *Face * Voice ID * Age* term, the model had to be evaluated for possible effects of *Face* and *Voice ID* at the 1st, 2nd, and 3rd quartile values of *Age*. Upon comparing the various *Face–Voice ID* combinations at the three *Age* quartiles, there appeared to be evidence of a congruency effect among the middle-aged and older listeners, but this effect seems to be specific to Foreign2; those who had heard Foreign2 paired with a congruent Moroccan face rated the voice as significantly less accented compared to those who had heard the same voice paired with an incongruent White face, or with no face at all. Since there were no other significant differences observed at the three *Age* quartiles, Table 5.9 and the interaction plot in Figure 5.6 below show the mean accentedness ratings for each *Face–Voice* pairing in a single *Age* quartile (= 53). The tables for the two other quartiles (*Age* = 23) and (*Age* = 31) can be found in Appendix M. Error bars represent standard errors of the mean.

Table 5.9 **Estimated mean accentedness ratings for each *Face–Voice* combination at *Age* = 53**

					95% Confidence Interval		
Face	Voice ID	Mean	SE	df	Lower	Upper	Group*
Audio	Native2	0.988	0.306	320	0.387	1.59	1
Audio	Native1	1.02	0.306	320	0.418	1.62	1
Moroccan	Native1	1.038	0.253	324	0.541	1.54	1
White	Native1	1.102	0.190	323	0.728	1.48	1
White	Native2	1.151	0.262	324	0.635	1.67	1
Moroccan	Native2	1.161	0.188	323	0.791	1.53	1
Moroccan	Foreign2	4.542	0.214	324	4.121	4.96	2
White	Foreign2	5.908	0.217	324	5.482	6.33	3
Moroccan	Foreign1	6.131	0.221	324	5.696	6.57	3
White	Foreign1	6.233	0.210	324	5.820	6.65	3
Audio	Foreign1	6.323	0.306	320	5.721	6.92	3
Audio	Foreign2	6.842	0.306	320	6.240	7.44	3

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

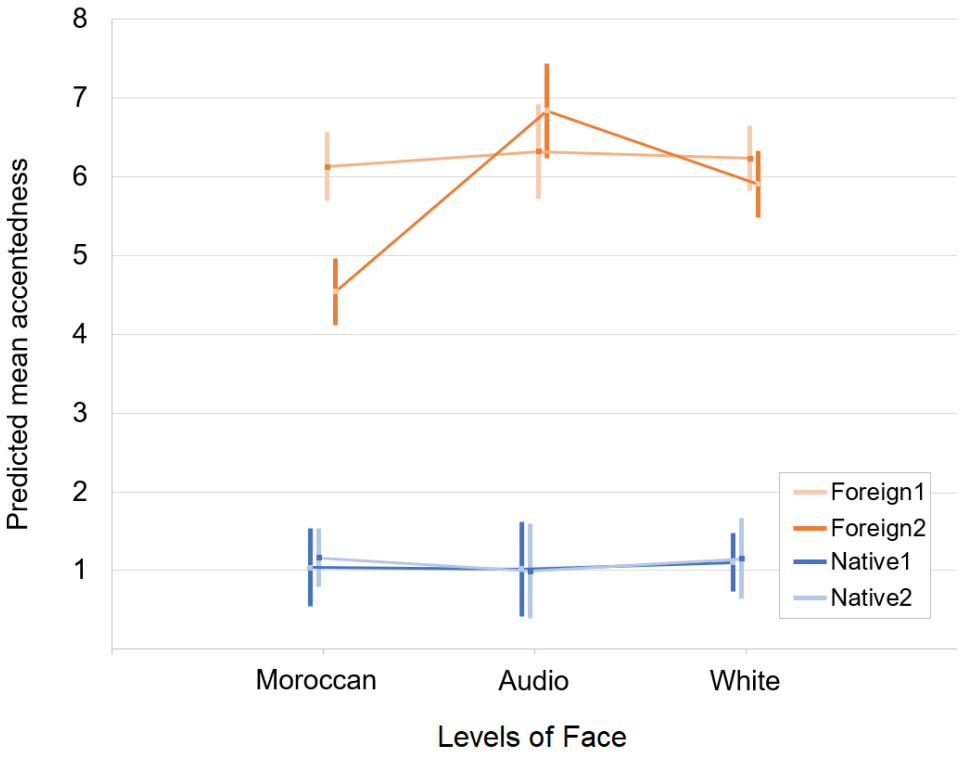


Figure 5.6 **Interaction plot of model-estimated mean accentedness scores for each *Face–Voice* combination at *Age* = 53**

As expected, both Dutch voices received scores close to 1 ('Dutch-sounding) on the seven-point scale, while the Moroccan voices were given significantly higher ('foreign-

sounding') ratings, with M1 receiving the highest accentedness ratings. The emmeans package was used to generate Table 5.10 below, which shows the overall model-estimated mean accentedness scores for each *Voice ID* when the accentedness scores were averaged across the *Face* and *Age* terms. As indicated in the "Group" column, the two SD speakers received significantly different accentedness ratings from the two MFD speakers, with Foreign1 being rated as the least native-sounding of the two MFD speakers, while the two Dutch voices were rated as mostly Dutch-sounding.

Table 5.10 Model-estimated mean accentedness ratings for each *Voice ID*

				95% Confidence Interval		
Voice ID	Mean	SE	df	Lower	Upper	Group*
Native2	1.12	0.096	321	0.926	1.304	1
Native1	1.19	0.097	321	1.002	1.384	1
Foreign2	5.86	0.096	319	5.668	6.047	2
Foreign1	6.33	0.096	320	6.142	6.519	3

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Next, just as in Chapter 4, contrasts were created to formally test for evidence of a congruency effect at each *Age* quartile. The first contrast shown in Table 5.11 below tested for an overall congruency effect. Looking at the table, there appears to be an overall congruency effect among the middle-aged (*Age* = 31) and older (*Age* = 53) participants, as the middle-aged and older participants gave higher accentedness ratings to incongruent guises compared to congruent ones.

Table 5.11 Overall congruency effect on accentedness ratings at the three quartiles of *Age*

				95% Confidence Interval			
Age quartile	Contrast ratio	SE	df	Lower	Upper	*t* ratio	*p*
Age = 23	0.210	0.150	323	-0.085	0.505	1.398	0.163
Age = 31	0.248	0.120	324	0.013	0.483	2.077	0.039
Age = 53	0.353	0.158	318	0.042	0.665	2.232	0.026

The next three tables show individual incongruency effect contrasts for each of the four voices at the three *Age* quartiles using a Bonferroni correction. As with the overall contrast, there appeared to be a congruency effect among only the middle-aged and older participants, and the incongruency effect appears to be specific to Foreign2, as this is the only voice whose accentedness ratings were different depending on congruency. In

summary, the three tables below demonstrate an incongruency effect for Foreign2 among the middle-aged and older participants.

Table 5.12 Congruency effect on accentedness for each *Voice ID* at *Age* = 23

				95% Confidence Interval			
Voice ID	Contrast Ratio	SE	df	Lower	Upper	*t* ratio	*p*
Native1	0.167	0.298	322	-0.5811	0.916	0.562	1.000
Native2	0.205	0.297	322	-0.5422	0.952	0.689	1.000
Foreign1	-0.192	0.296	322	-0.9341	0.551	-0.648	1.000
Foreign2	0.659	0.300	322	-0.0936	1.412	2.199	0.114

Table 5.13 Congruency effect on accentedness for each *Voice ID* at *Age* = 31

				95% Confidence Interval			
Voice ID	Contrast Ratio	SE	df	Lower	Upper	*t* ratio	*p*
Native1	0.106	0.239	322	-0.495	0.706	0.441	1.000
Native2	0.153	0.236	322	-0.440	0.745	0.648	1.000
Foreign1	-0.113	0.236	322	-0.706	0.479	-0.481	1.000
Foreign2	0.848	0.238	322	0.249	1.446	3.559	0.002

Table 5.14 Congruency effect on accentedness for each *Voice ID* at *Age* = 53

				95% Confidence Interval			
Voice ID	Contrast Ratio	SE	df	Lower	Upper	*t* ratio	*p*
Native1	-0.064	0.316	322	-0.859	0.731	-0.203	1.000
Native2	0.010	0.323	322	-0.801	0.821	0.031	1.000
Foreign1	0.102	0.305	322	-0.665	0.868	0.333	1.000
Foreign2	1.366	0.305	322	0.601	2.132	4.483	< .0001

Discussion Analysis II

As expected on the basis of previous research, this analysis revealed that the SD speakers received significantly lower accentedness ratings than the MFD speakers, i.e., the SD speakers were rated as more 'Dutch-sounding,' while the MFD speakers were rated as more 'foreign-sounding.' Yet contrary to the null finding reported in Chapter 4, participant guise did appear to differentially affect the accentedness ratings of a single MFD speaker (Foreign2), as the incongruent pairing of this voice with a White face entailed significantly higher accentedness scores among middle-aged and older participants than

when the MFD voice had been shown belonging to a Moroccan face. There was no such incongruency effect observed among the younger listeners. Because this congruency effect was found only for Foreign2, and not among all listeners, this finding cannot be interpreted as strong evidence in favor of the exemplar theory. Nonetheless, the observation that congruency did appear to play a role in the accentedness rating of Foreign2 is the closest this experiment has come to supporting the predictions of the exemplar theory.

5.5.3. Analysis III: Credibility scores

This final analysis tested whether specific voices, faces, or a combination thereof would elicit different credibility ratings. The scores (1 = definitely true, 7 = definitely false) were fitted to several linear mixed effect models, which will each be further explained in the sections below. All analyses included *Participant, SV* (a term accounting for a possible interaction effect between *Voice ID* and *Sentence*) and *Sentence* as random effects. Since the assumptions of the analysis were not violated (as determined by a Q-Q plot), the data did not require transformation.

All ratings included

A linear mixed effects model was fit to the credibility scores, with the same fixed effects *Face, Voice,* and *Veracity* as in the response time analysis that was described in Section 5.5.1. An ANOVA was carried out on the fitted model, and as can be seen in the ANOVA results below, this analysis yielded no evidence of visual speaker ethnicity (*Face*), accent (*Voice),* sentence veracity, or participant age differentially affecting the various guises' perceived believability. In other words, the various guises did not elicit different credibility ratings among the listeners.

Table 5.15 Type III Analysis of Variance Table with Kenward-Roger's method

Fixed effect	Sum of squares	Mean squares	Num DF	Den DF	*F*	*p*
Voice ID	7.050	2.350	3	125	0.967	0.411
Face	4.201	2.100	2	216	0.862	0.424
Veracity	3.834	3.834	1	41	1.578	0.216
Voice ID * Face	5.294	0.882	6	2539	0.363	0.903
Voice ID & Veracity	5.242	1.747	3	125	0.719	0.542
Face * Veracity	0.858	0.429	2	3558	0.176	0.838
Voice ID * Face * Veracity	6.618	1.103	6	3565	0.454	0.843

Subsequent post hoc tests did not provide any evidence of *Face* or *Voice* effects either. Since there were no significant differences, the estimated mean credibility scores for each level of *Voice ID,* each level of *Face*, and each *Face–Voice ID* combination can be found in Appendix P. Contrasts furthermore yielded no indication of incongruency effects (Table 5.16), or AV effects (Table 5.17) in this data.

Table 5.16 Contrast of mean credibility scores for incongruent vs. congruent *Face–Voice* combinations

			95% Confidence Interval			
Contrast Ratio	SE	df	Lower	Upper	*t* ratio	*p*
-0.021	0.066	1772	-0.149	0.108	-0.314	0.754

Table 5.17 Contrast of mean credibility scores for AV vs. AO conditions

			95% Confidence Interval			
Contrast Ratio	SE	df	Lower	Upper	*t* ratio	*p*
0.0722	0.081	83	-0.088	0.233	0.896	0.373

Confidence-coded credibility ratings

Rather than using a dichotomous ‘true/false’ scale to determine the believability of the trivia sentences, this experiment used a seven-point scale, which allowed the participants to simultaneously indicate how confident they were about their judgment. The two far ends of the scale (i.e., 1 and 7) were used to indicate the highest levels of confidence, while the two numbers closest to the middle point (i.e., 3 and 5) indicated the highest degree of doubt. In order to test whether certain guises entailed more confident

answers than others, participant responses were recoded according to how far they were removed from the neutral middle point of the rating scale. See Table 5.18 for a breakdown of the recoded and color-coded scores.

Table 5.18 Recoded credibility scores

Original score	Descriptor	Recoded score
1	Definitely true	3
2	True	2
3	Probably true	1
4		0
5	Probably false	1
6	False	2
7	Definitely false	3

The same linear mixed effects model with the same random effects was used. The ANOVA that was carried out on this fitted model (see Table 5.19) revealed no evidence of any *Face, Voice ID, Age,* or sentence *Veracity* effects on the participants' confidence levels, nor did subsequent tests reveal any significant differences. This meant that participants were equally confident in their responses across the four guises.

Table 5.19 Type III Analysis of Variance Table with Kenward-Roger's method

Fixed effect	Sum of squares	Mean squares	NumDF	DenDF	*F*	*p*
Voice ID	2.997	0.999	3	124	1.742	0.162
Veracity	0.245	0.245	1	41	0.427	0.517
Face	1.455	0.728	2	218	1.265	0.284
Voice ID * Veracity	1.058	0.353	3	124	0.615	0.607
Voice ID * Face	5.584	0.931	6	3510	1.623	0.137
Veracity * Face	0.271	0.135	2	3558	0.236	0.790
Voice ID * Veracity * Face	4.666	0.778	6	3562	1.356	0.229

Certainties removed

The premise of the accent credibility assessment task was to test whether listeners who were unaware of the correct answer to a trivia statement (and therefore could go either way) would be influenced in their veracity judgment by the person saying it. Since participants had not been given the option to separately indicate when they already knew the veracity of the trivia statement, it became difficult to make a distinction between confident guesses and responses based on actual prior knowledge. This methodological

oversight made it unclear to what extent the guises had actually influenced the judgments on the two far ends of the scale. However, it is worth pointing out that Lev-Ari and Keysar (2010) found that listeners' prior knowledge did not seem to prevent them from rating the foreign-accented speakers as less believable anyway.

By removing the most confident ratings on the two far ends of the rating scale from the dataset (i.e., 1 and 7), this analysis attempted to remove responses that could possibly have been informed by prior knowledge. However, after re-running the same model as was described in 5.5.1, no evidence was found of visual speaker ethnicity, accent, or incongruency differentially affecting listeners' credibility judgments of the trivia sentences.

Table 5.20 Type III Analysis of Variance Table with Kenward-Roger's method

Fixed effect	Sum of squares	Mean squares	NumDF	DenDF	*F*	*p*
Voice ID	2.082	0.694	3	121	0.414	0.743
Veracity	7.544	7.544	1	41	4.498	0.040
Face	1.427	0.714	2	207	0.424	0.655
Voice ID * Veracity	4.498	1.499	3	121	0.894	0.446
Voice ID * Face	3.742	0.624	6	1832	0.372	0.897
Veracity * Face	0.799	0.400	2	2867	0.238	0.788
Voice ID * Veracity * Face	6.232	1.039	6	2870	0.619	0.715

Since the ANOVA indicated a significant main effect of sentence *Veracity* on the credibility ratings, the emmeans package was used to generate the model-estimated credibility means for the two types of sentences. As can be seen in Table 5.21 below, the false sentences received significantly higher credibility ratings than the true ones.

Table 5.21 Estimated mean credibility ratings of true and false statements

				95% Confidence Interval	
Veracity	Mean	SE	df	Lower	Upper
False	3.63	0.121	42	3.387	3.874
True	3.99	0.124	42	3.745	4.244

Discussion Analysis III

This experiment's credibility ratings did not support Lev-Ari and Keysar's (2010) finding that foreign-accented speech is rated as less believable than native speech, since the guises speaking in an MFD voice were not perceived as any less credible than guises

speaking in a SD voice. This was in line with what was predicted on the basis of previous studies' failure to replicate Lev-Ari and Keysar's (2010) finding (De Meo, 2012; De Meo et al., 2011; Frances et al., 2018; Souza & Markman, 2013; Stocker, 2017). These results thus contest Lev-Ari and Keysar's (2010) conclusion that voices that are more difficult to process are perceived as less credible, as the MFD speakers in this study did have longer response times, suggesting lower comprehensibility than the SD speakers. Interestingly, when the most confident ratings were removed from the dataset, the true sentences were rated as significantly less credible than the false sentences, but since this difference was very small (0.36 on a 5-point scale), no strong conclusions can be drawn about the implications of this finding.

Although this experiment has failed to replicate Lev-Ari and Keysar's (2010) finding, it must be noted that this experiment used a slightly different approach, as it was conducted on native Dutch listeners as opposed to native English listeners, and it featured speakers of an ethnolect rather than a foreign accent. It is nonetheless unlikely that this change in methodology was responsible for the observed null effect, since similar studies focusing on the credibility of foreign accents on native listeners also failed to replicate these findings, just as regional accents were not found to be any less credible than local accents (Frances et al., 2018). Yet perhaps the type of tasks used in these experiments did not truly tap into credibility after all. Rather than using trivia statements, future research could try to test credibility by measuring the extent to which native and non-native teachers explaining an obscure grammar point are believed. This shift away from trivia statements to a more real-life educational setting with a power differential between the interlocutors may result in different findings.

Because there was no difference in credibility observed between the various guises, this experiment showed no evidence that visual speaker ethnicity influenced guise credibility. This is valuable information, as this had not been formally tested in previous work. The other two analyses furthermore demonstrated that all four guises entailed roughly equally confident answers, and that there were also no differences in credibility ratings between the four guises when prior knowledge was controlled for. Instead, participants were just as likely to judge a trivia statement as true or false–regardless of which guise had said it.

5.6. Discussion of Experiment 3

This experiment was designed to replicate the response time and accentedness findings reported in Chapter 4, as well as test Lev-Ari and Keysar's (2010) conclusion that non-native speakers are perceived as less credible than native speakers. In contrast with Experiment 2 (which had used White Canadian and Japanese native speakers), this experiment featured Dutch and Moroccan women speaking in a Standard Dutch dialect and in an ethnic dialect (Moroccan-flavored Dutch). The expectation was that the different cultural context and participants' attitudes towards Moroccan Dutch speakers could potentially elicit different responses than had been observed in Vancouver. Yet because of conflicting past results, no predictions were made about the effect of visual speaker ethnicity on accentedness ratings and processing speeds.

Consistent with the findings of Experiment 2, comprehensibility appeared to be unaffected by visual speaker ethnicity in any way; similarly, participant age was not found to differentially affect response times. Because both exemplar models and the reverse linguistic stereotyping hypothesis assume some effect of visual speaker information, this experiment's response time data did not provide support for either theoretical framework. Certain accentedness ratings did appear to partially support the exemplar theory, though, as there seemed to be a congruency effect for a specific MFD speaker's voice among the middle-aged and older participants. Listeners who had seen a Moroccan guise speaking with the congruent voice Foreign2 assigned significantly lower accentedness ratings to that voice than when that same voice had been paired with an incongruent White guise or in the Audio-Only condition. The accentedness rating data furthermore supported the experiment's hypothesis that the MFD speakers would receive significantly higher (i.e., more 'foreign-sounding') accentedness ratings than the SD speakers–regardless of guise ethnicity. Yet contrary to expectation, accentedness ratings were no different across the three age quartiles, i.e., there were no differences in accentedness ratings between the young, middle-aged, and older listeners. Note that in this experiment, listeners were reminded what each of the guises had sounded like before assigning accentedness ratings, whereas Experiment 2 relied more heavily on listeners' memory, since they were asked to provide ratings for all four voices at the end of the experiment without hearing each voice once more.

Although this experiment replaced the foreign accent in Experiments 1 and 2 with an ethnolect, the data accord with our earlier observations and with Munro and Derwing (1995). Here too, MFD speech required more time to respond to than SD speech, thereby confirming that MFD is more difficult to process. It is however important to acknowledge that the voice effects reported in this chapter may be specific to the four individual speakers featured in this experiment, so we cannot simply generalize this observed difference to all MFD and SD speakers.

Finally, the results from the accent credibility assessment task were inconsistent with Lev-Ari and Keysar's (2010) conclusion that increased processing difficulty negatively affects speaker credibility. Although MFD is an ethnolect rather than a foreign dialect, this experiment established that MFD voices took significantly longer to process than SD voices. According to Lev-Ari and Keysar (2010), then, the lower comprehensibility of the MFD voices should have negatively impacted their credibility, but this proved not to be the case. In fact, listeners were equally likely to judge a trivia sentence as true or false, regardless of what the guise sounded *or* looked like, i.e., visual speaker ethnicity was not found to influence perceived credibility either.

Chapter 6.

General discussion and conclusions

6.1. Summary of results

Considerable research to date has tended to focus on whether listeners' social expectations and racial biases differently affect the intelligibility of speakers and reported accentedness ratings. Since no study had examined whether visual speaker ethnicity also influences comprehensibility, the main aim of the three response time studies described in this dissertation was to formally test the effect of a speaker's ethnic appearance on speech processing speeds. A secondary aim of this research was to test whether the recorded response times, accentedness, and intelligibility data would provide support for either the exemplar theory or the reverse linguistic stereotyping framework.

The first response time task presented **Chapter 3** used photographs of an Asian and a White speaker paired with two Canadian-English and two Japanese-accented voices to test whether the processing speeds of sentences would be influenced by guise ethnicity. However, the methodological decision to pair the same two photographs with all four voices complicated the interpretation of the effect of visual speaker ethnicity, as it was unclear whether listeners had even attended to visual speaker ethnicity, or instead had considered it a distraction. As such, the lack of an ethnicity effect cannot be interpreted as convincing evidence against an ethnicity effect. Results did reveal that the native English listeners assigned higher accentedness ratings to the Japanese speakers than to the Canadian speakers, and that it took them significantly longer to process foreign-accented speech compared to native speech. Listeners were also more likely to make sentence identification errors when listening to Japanese-accented statements. Finally, low comprehensibility ratings were associated with longer response times, just as Munro and Derwing (1995) found.

The second response time task in **Chapter 4** presented listeners with dubbed videos of two Asian and two White guises to decrease demand characteristics. This time the face–voice pairings were kept constant for each participant throughout the experiment, and sentences were presented under degraded listening conditions (multi-talker-babble).

Still, visual speaker ethnicity was not found to influence processing speeds. The Japanese speakers did elicit longer response times and higher accentedness ratings than their native English-speaking counterparts, thereby corroborating the findings reported in both Experiment 1 and Munro and Derwing (1995). Results from a speech shadowing task furthermore showed that listeners had more difficulty correctly repeating the Japanese-accented voices compared to the Canadian-accented ones, but that the speakers' purported ethnicity did not affect these intelligibility scores.

The third reaction time task presented in **Chapter 5** was conducted in the Netherlands and used videos of White and Moroccan women speaking Standard Dutch and Moroccan-flavored Dutch. Just as in Experiment 2, sentences were embedded in multi-talker-babble. Although this experiment was conducted in a different cultural setting and in a different language from the other two experiments, the results were similar, in that it yielded no evidence of visual speaker ethnicity influencing listeners' responses times. Like the previous two experiments, accent was once again found to play a role, as the two Moroccan-flavored Dutch voices took longer to process and were rated as more accented compared to the two Standard Dutch-speaking voices. This experiment also found some evidence of visual speaker ethnicity affecting accentedness ratings, but only for a single Moroccan-flavored Dutch speaker; when that voice was paired with an incongruent White face, some of the listeners rated the voice as more accented than when it was paired with a congruent Moroccan face. Contrary to prediction, listener age was not found to differentially affect response times and accentedness ratings.

The results from the accent credibility assessment task proved to be inconsistent with Lev-Ari and Keysar's (2010) conclusion that increased processing difficulty negatively affects speaker credibility, as there were no differences in believability between the Moroccan-flavored Dutch and Standard Dutch voices–even though the Moroccan-flavored Dutch speakers had lower comprehensibility. The ethnicity of the speaker did not appear to affect credibility judgements either, as the trivia sentences were equally likely to be judged as true or false, regardless of guise.

6.2. Implications

The first two experiments described in this dissertation both successfully replicated Munro and Derwing's (1995) finding that non-native-sounding voices take longer to process than native-sounding voices–irrespective of visual speaker ethnicity. Experiment 3 furthermore showed that non-standard speech also elicited longer reaction times than standard speech. Assuming that reaction times reflect comprehensibility (Munro & Derwing, 1995), these findings provide additional support for the hypothesis that both foreign-accented speech and non-standard speech have lower comprehensibility than native-accented or standard speech. Experiment 1 furthermore substantiated Munro and Derwing's (1995) observation that low comprehensibility ratings were associated with longer response times, while Experiment 2 revealed that listeners were slightly worse at shadowing foreign-accented voices, which suggests lower intelligibility. Finally, the finding that the veracity of non-native-sounding utterances (and non-standard-sounding utterances in Experiment 3) was more often incorrectly identified emphasizes the fact that listeners find it more difficult to interpret foreign-accented or non-standard speech compared to utterances produced in their own dialect.

While previous research findings suggest that visual speaker ethnicity influences accentedness ratings and intelligibility scores (e.g., Babel & Russell, 2015; McGowan, 2011), this project was unable to replicate these findings with the exception of a single significant difference in accentedness ratings that appeared to be motivated by face–voice incongruency (see section 5.5.2). As such, the results failed to provide compelling support for either the exemplar theory or the reverse linguistic stereotyping hypothesis, as both predict that visual speaker information in combination with auditory input would differentially affect how speech would be evaluated. Furthermore, none of the response time tasks provided evidence that visual speaker ethnicity influenced comprehensibility. The overall implication of this finding seems to be that visual speaker ethnicity may not meaningfully impact listeners'–largely automatic–speech processing. Section 6.2.1. will present a number of possible reasons for the lack of an ethnicity effect on response times and the disparity in the results of my experiments compared to previous findings.

Experiment 3 featured the first accent credibility assessment task to investigate whether trivia statements spoken in a Dutch ethnolect (Moroccan-flavored Dutch) were rated as less believable than when they were spoken in a Standard Dutch accent. Results

demonstrated that although it took listeners significantly longer to process Moroccan-flavored Dutch speech, the Moroccan-flavored Dutch voices were rated as equally credible compared to the Standard Dutch voices. These findings thus run counter to Lev-Ari and Keysar's (2010) inference that increased processing difficulty negatively affects speaker credibility. Experiment 3 was also the first experiment to formally test whether visual speaker ethnicity could affect credibility, but since no such effect was observed, the results provided additional evidence for the inconsistent effect of visual speaker ethnicity on speaker evaluations. The overall implication of these null effects, in conjunction with four similarly 'failed' replications, is that linguists should reconsider the validity of the claim that having a foreign accent "can unfairly destroy your credibility" (M. Schmid, 2019), seeing as the contradictory available data is inconclusive as to the relationship between accent and credibility. It may be advisable to convey these new findings to the wider public, because Lev-Ari and Keysar's (2010) findings received a lot of attention in the media.

6.2.1. Comparison to previous research

This dissertation has critically evaluated previous research to determine how some of the contradictory findings in previous work may have come about. Since none of the studies were identical in design, it is difficult to determine which factors may have contributed the conflicting past results. Nonetheless, a list of potential reasons is provided below.

Study design

One of the major contributing factors to observing an effect of visual speaker ethnicity appears to be study design. This suggestion was also made by Hanulíková (2018), who posited that "it is possible that effects of ethnicity on language processing are mitigated when the two speakers are presented one after another and might bear out more strongly depending on the design and task" (p. 7). As indicated earlier, Zheng and Samuel (2017) were able to manipulate their results by changing the study design. When they used the same between-subjects design as Rubin (1992), i.e., a single accentedness rating based on a single picture paired with a voice, they also found that the Asian guises were rated as more accented than the White guises. Yet when listeners were asked to rate the other, remaining guise as well (so they had now rated *both* the White and Asian guise), this pattern reversed, thereby highlighting the effect's apparent high sensitivity to

response strategies when pictures were used (p. 1847). Indeed, the way stimuli are presented seems to influence findings significantly; although Zheng and Samuel (2017) found a small effect of visual speaker ethnicity when they used videos blocked on ethnicity, this effect disappeared when they presented the same videos in a randomized order.

Different Stimuli

Using photographs instead of videos also seems to elicit different results; all of the studies that used photographs observed some sort of 'ethnicity effect,' with different guises receiving dissimilar accentedness ratings and intelligibility scores (i.e., Babel & Russell, 2015; Hanulíková, 2018; Kang & Rubin, 2009; McGowan, 2015; Rubin, 1992; Rubin et al., 1999, 2016; Zheng & Samuel, 2017). Yet among the studies that used videos instead, the results were much less straightforward. Some also reported finding an ethnicity effect (i.e., Babel & Mellesmoen, 2019; Gnevsheva, 2018; Zheng & Samuel, 2017), while others–including the experiments described in Chapters 4 and 5 of this dissertation–did not (i.e., Karpinska, 2019; Yi et al., 2014; Zheng & Samuel, 2017).[33]

Statistical methods and analysis

Whether a guise ethnicity effect is observed also appears to be dependent on researchers' statistical methods and how the results are interpreted. While studies conducted pre-2000 tended to use ANOVAs (which required averaging over many datapoints), more recent psycholinguistic and sociophonetic studies, including this one, have increasingly relied on sophisticated statistical models that not only retain all data points, but also enable researchers to take many more factors into account (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). Additionally, many of the discussed studies that used statistical models to analyze their data did not report effect sizes in their analyses (e.g., Babel & Russell, 2015; Gnevsheva, 2018; Hanulíková, 2018; Karpinska, 2019; McGowan, 2015; Yi et al., 2014), which makes comparisons across studies difficult (see Yi et al. [2013] for such an example). Researchers furthermore use varying criteria for defining outliers. The decision to keep specific datapoints or to remove them from the dataset can make the difference between a significant and non-significant effect, so the influence that outliers may have on study outcomes should not be underestimated. Future

[33] I put Yi et al (2013) in the category of studies that did not find an ethnicity effect, because the actual difference was too minor to be meaningful –especially since there was no effect size given.

research will have to determine whether these different statistical approaches may have contributed to the contradictory findings.

Offline vs. online tasks

A final observation is that, with the exception of Staum Casasanto (2008, 2010) and Squires (2013), all the experiments focusing on the effect of visual speaker ethnicity used offline tasks in which listeners were able to reflect on their answers and even change them. Consequently, –although interesting and informative in their own way–these types of tasks do not necessarily provide compelling insights into listeners' perceptual processes. Rather, they appear to address the question of how the more agentive interpretative level is influenced by this type of speaker information (Firestone & Scholl, 2016; Zheng & Samuel, 2017). Nonetheless, most researchers have interpreted their findings as constituting evidence of listeners' speech *perception* being influenced. Even though the current research did not find a convincing effect of visual speaker ethnicity in its semi-online and offline tasks, it remains beyond the scope of this work to suggest that the findings provide supporting evidence for Zheng and Samuel's (2017) submission that speech perception may be entirely separate from post-stimulus interpretation.

In sum, in this dissertation I have highlighted the apparent labile nature of the purported effect of visual speaker ethnicity on speech processing and evaluation. The null findings described here challenge the conclusions drawn in previous research, which suggest that there is evidence of racial bias or an expectation mismatch effect. In principle, if an expectation mismatch effect or reverse linguistic stereotyping actually exists, variation in study design, stimuli, tasks, or type of analysis should have a minimal impact on outcomes. Therefore, the contradictory observations in the existing literature should encourage researchers to re-evaluate some of the findings that are currently considered seminal.

6.3. Limitations

This research has several limitations. Since the experiments featured only two speakers per accent, the findings are likely not generalizable to the overall accent categories investigated in the experiments. Based on the limited number of speakers for each experiment, it is important not to conflate talker-specific features with accent-specific

features, as there is a real possibility that some of the observed differences among speakers were idiosyncratic rather than accent-related. This challenge becomes evident in the considerable variation between speakers within the same accent category; one of the two Japanese speakers in Experiment 2, for instance, proved to be significantly less intelligible than the other, while the two Moroccan-flavored Dutch voices in Experiment 3 entailed significantly different accentedness ratings and response times from one another. Furthermore, the occasionally unnatural-sounding intonational patterns of the Moroccan-flavored speakers (as opposed to the natural-sounding Standard Dutch renditions) may have affected RTs differently as well. Another limitation to this research is that speaker faces were grouped together into ethnic categories to keep the number of levels in the analysis manageable, whereas the various voices were analyzed separately. This methodological decision obscured any idiosyncratic differences between the speakers' faces, and therefore may have misattributed the effect of idiosyncratic features to overall speaker ethnicity. The implications of these findings are furthermore restricted to native listeners, as non-native listeners were deliberately excluded from the experiments to (1) make the results more directly comparable to previous research, and (2) to keep the number of independent variables manageable.

In addition to the above-mentioned limitations to generalizability, the experiments' findings may have been influenced by a number of response biases. Even though most listeners did not realize that the speakers' ethnicities and accents had been manipulated, the very fact that each experiment featured two White and two visible minority speakers may have given them a sense of the purpose of the experiment. As a result, demand characteristics may have influenced participant responses, particularly in the offline tasks in which listeners had more time to think about their answers. The results of the offline rating tasks must also be considered in light of their sensitivity to social desirability bias, since each experiment featured both White and non-White guises, thereby possibly making listeners aware that visual speaker ethnicity may play a role. Even though some people are willing to judge others based on their accents (Munro, 2003), they may be a bit more reluctant to overtly evaluate an individual based on their ethnic or racial background. It is therefore possible that listeners (sub)consciously changed their offline accentedness and credibility assessment ratings to more socially appropriate ones when rating the various ethnically-different guises.

Some unexpected challenges emerged during the research process. Most notably, the methodological decision to show the same two photographs paired with four different voices in Experiment 1 had the unintended consequence that a speaker's purported ethnicity was not consistent throughout the experiment, which in turn led the listeners to regard the speaker's purported ethnicity as a distraction rather than as useful speaker information. This meant that the findings of Experiment 1 must be interpreted with caution, as the observed null effect of guise ethnicity on RTs, correctness scores, and accentedness ratings may simply reflect the listeners' disregard of the visual information. This methodological complication was corrected in Experiments 2 and 3, where the face–voice combinations were kept constant throughout for each participant.

Another unexpected methodological challenge was that the stimulus presentation software used in Experiment 2 had increasing difficulty retrieving the video files over Wi-Fi, which likely triggered the steadily increasing cumulative differences in response times between the AV and AO conditions as the experiment progressed. Nevertheless, since this software malfunction affected comparisons only between the AV and AO conditions, Experiments 2 and 3 still provide a valuable insight into the effect of visual speaker ethnicity on speech processing.

Finally, the survey that was used in Experiment 1 to collect data on listener exposure to Japanese-accented English revealed some of the challenges with reliably quantifying listeners' previous experiences. For instance, Experiment 1 featured the question 'how often would you say you interact with a Japanese-accented speaker? (daily, weekly, monthly, almost never)' to which one participant responded that he did not know how to answer this question as he had lived in Japan for nearly two years but had since moving back and had not encountered Japanese-accented speakers very often. Based on this feedback and my subsequent preference not to divide participants into a dichotomous variable of 'more experienced' and 'less experienced' listeners based on potentially ambiguous background questions, the decision was made not to include measurements of listener exposure to Japanese-accented English and Moroccan-flavored Dutch in the other two experiments. This methodological decision did mean that potentially valuable background information was left out of the subsequent two analyses.

6.4. Future directions

Use both online and offline measures

Most of the studies that cited evidence of visual speaker ethnicity affecting perception, with the exception of D'Onofrio (2015) and Koops et al. (2008), used offline measures such as gap texts and transcription tasks. While these methodological decisions are sound in and of themselves, the lack of variety in experimental methodologies has considerably limited the field's focus and scope. Further research is therefore needed to establish whether the effect of visual speaker ethnicity on accentedness ratings, intelligibility, and comprehensibility can also be observed in online measures. Online methods making use of eye-tracking, ERPs and EEG will undoubtedly provide invaluable insights to the field, as they have remained underused when investigating this phenomenon due to their prohibitive costs.

Use non-native listeners

Most studies investigating the effect of visual speaker ethnicity on listeners have been carried out in the North American context, specifically on native listeners of English. So far, Karpinska (2019) is the only researcher to have tested the effect of visual speaker ethnicity on non-native listeners.[34] This focus on native listeners has considerably limited the generalizability of the reported findings; future research on the effect of visual speaker information would benefit from focusing on, or at the very least including, non-native listeners in their designs. Additionally, in view of Tekin's (2019) finding of an effect of a listener's ethnic group affiliation on accentedness ratings, future studies should attempt to recruit listeners with varied linguistic and ethnic backgrounds.

Static vs. fluid social categories

Finally, future studies should take personae into account when determining the effect of social identity on speech evaluation. D'Onofrio (2019) discovered that various photographs of the same man elicited different results, depending on what listeners inferred his 'persona' to be (which was manipulated by changing the individual's hairstyle,

[34] Although Kang and Rubin's (2009) experiment featured both non-native and native listeners, they did not analyze the interaction between the nativeness of the listeners and their other dependent variables.

pose, and clothing style). This observation underlines the fluidity of the concept of ethnicity, and simultaneously casts some doubt on what studies using photographs were actually testing, since the way the purported guises looked, dressed, and posed in the photo may have influenced how listeners evaluated them, rather than their ethnicity. Additionally, D'Onofrio's (2019) findings also highlight the potential influence that listeners' perceived personalities may have on their ratings.

Sociophonetics is still an emerging research field. The fact that this research is the first to survey the relationship between visual speaker ethnicity and processing speeds testifies to this. Aside from replicating the experiments described in this dissertation with the above recommendations in place, replication studies in this line of research will contribute meaningfully to the field and have the potential to clarify some of the perplexing and contradictory observations reported in the currently limited body of research. To use the concluding statement of a large-scale replication study of 100 psychology studies: "innovation points out paths that are possible; replication points out paths that are likely; progress relies on both" (Open Science Collaboration, 2015, p. 4716-7).

6.5. Conclusion

In response to prior research reporting an effect of visual speaker ethnicity on accentedness ratings and transcription accuracy, this research is the first to formally assess whether a person's ethnicity also exerts an influence on listeners' processing speeds. Based on the results from three response time studies, it appears that comprehensibility was unaffected by this type of social information. Since the results from the accentedness rating tasks and speech shadowing task similarly failed to confirm an overall effect of visual speaker ethnicity, the findings from this research did not provide compelling evidence to support the predictions of either the exemplar theory or the reverse linguistic stereotyping hypothesis, as both theoretical frameworks presuppose an effect of visual speaker ethnicity on speech processing/evaluation. The observation in the third experiment that congruency did appear to play a role when rating the accentedness of a specific foreign-accented voice is the closest this study came to supporting the predictions of the exemplar theory. Yet because the finding was restricted to a specific voice and only found in two of the three *Age* groups, this finding should not be considered compelling evidence.

In addition to replicating the findings of Munro and Derwing (1995), this dissertation also critically examined the methodologies and reported findings of prior relevant sociophonetic work. The main conclusion drawn is that, due to the nature of the tasks used in previous research, there is no compelling evidence to demonstrate that listener *perception* is influenced by visual speaker ethnicity. In fact, the lack of consistent evidence supporting this premise has led me to propose that the phenomenon may be a methodological artefact–especially since the effect has been shown to be sensitive to methodological choices (see Zheng & Samuel, 2017). To test this proposition, many further studies on the topic are required.

Replication studies relating to the themes covered in this dissertation should ideally contrast different methodological approaches to determine whether phenomena can be observed in a range of experiments employing diverse designs. For example, they might compare responses to the same stimuli in a within-subjects vs. between-subjects design or investigate how presentation order affects the results (e.g., by using a blocked design vs. a mixed design, or employing adaptors).

In sum, this study has highlighted the fact that without more research to shed light on what has caused the contradictory findings reported so far, researchers in this field should refrain from making expansive claims about the implications of their research findings, as this is not the type of subject matter that merits premature conclusions.

References

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 520–529. https://doi.org/10.1037/a0013552

Adobe Systems. (2019). *Adobe Premie Pro*. Retrieved from https://www.adobe.com/ca/products/premiere.html

Angus Reid Institute. (2018). *Immigration in Canada: Does recent change in forty year opinion trend signal a blip or a breaking point?* 1–11. Retrieved from http://angusreid.org/wp-content/uploads/2018/08/2018.08.01_Immigration-release.pdf

Arends-Tóth, J., & Vijver, F. J. R. Van De. (2003). Multiculturalism and acculturation: views of Dutch and Turkish-Dutch. *European Journal of Social Psychology*, *33*(2), 249–266. https://doi.org/10.1002/ejsp.143

Audacity Development Team. (2018). *Audacity: Free Audio Editor and Recorder*. Retrieved from https://sourceforge.net/projects/audacity/

Babel, M., & Mellesmoen, G. (2019). Perceptual adaptation to stereotyped accents in audio-visual speech. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia* (pp. 1044–1048). Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Babel, M., & Russell, J. (2015). Expectations and speech intelligibility. *The Journal of the Acoustical Society of America*, *137*(5), 2823–2833. https://doi.org/10.1121/1.4919317

Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015). Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation. *Frontiers in Human Neuroscience*, *9*(422), 1–13. https://doi.org/10.3389/fnhum.2015.00422

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bent, T., & Holt, R. F. (2017). Representation of speech variability. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(4), 1–14. https://doi.org/10.1002/wcs.1434

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A fieldeExperiment on labor market discrimination. *American Economic Review*, *94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

Bettencourt, B. A., Dill, K. E., Greathouse, S. A., Charlton, K., & Mulholland, A. (1997). Evaluations of Ingroup and Outgroup Members: The Role of Category-Based Expectancy Violation. *Journal of Experimental Social Psychology*, *33*(3), 244–275. https://doi.org/10.1006/jesp.1996.1323

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer (6.0.40)*. Retrieved from http://www.praat.org/

Bouavichith, D. A., Calloway, I., Craft, J. T., & Hildebrandt, T. (2019). Perceptual influences of social and linguistic priming are bidirectional. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia* (pp. 1039–1043). Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Bradac, J. J., Cargile, A. C., & Hallett, J. S. (2001). Language attitudes: Retrospect, conspect, and prospect. In W. P. Robinson & H. Giles (Eds.), *The new handbook of language and social psychology* (pp. 137–158). Chichester, UK: John Wiley.

Bradlow, A. R., & Bent, T. (2003). Listener adaptation to foreign-accented English. In M. J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain* (pp. 1581–1583). Universitat Autònoma de Barcelona.

Brown, K. (1992). American college student attitudes toward non-native instructors. *Multilingua: Journal of Cross-Cultural and Interlanguage Communication*, *11*(3), 249–266. https://doi.org/10.1515/mult.1992.11.3.249

Browne, R. H. (2010). The t-test p value and its relationship to the effect size and P(X >Y). *The American Statistician*, *64*(1), 30–33. https://doi.org/10.1198/tast.2010.08261

Brückner, S., & Kammer, T. (2017). P286 How tDCS modulates semantic processing: Online versus offline stimulation. *Clinical Neurophysiology*, *128*(3), e150–e150. https://doi.org/10.1016/j.clinph.2016.10.394

Bursell, M. (2007). What's in a name? A field experiment test for the existence of ethnic discrimination in the hiring process. *2007:7, SULCIS Working Papers*, 1–28.

Caffarra, S., & Martin, C. D. (2019). Not all errors are the same: ERP sensitivity to error typicality in foreign accented speech perception. *Cortex*, *116*, 308–320. https://doi.org/10.1016/j.cortex.2018.03.007

Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., & Rodd, J. M. (2017). Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition. *Cognitive Psychology*, *98*, 73–101. https://doi.org/10.1016/j.cogpsych.2017.08.003

Campbell-Kibler, K. (2009). The nature of sociolinguistic perception. *Language Variation and Change*, *21*(1), 135–156. https://doi.org/10.1017/S0954394509000052

Campbell-Kibler, K., & McCullough, E. A. (2015). Perceived foreign accent as a predicator of face-voice match. In A. Prikhodkine & D. R. Preston (Eds.), *Responses to language varieties: Variability, processes and outcomes* (pp. 176–189). John Benjamins Publishing Company.

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

Carreiras, M., & Clifton, C. E. (Eds.). (2004). *The on-line study of sentence comprehension: Eyetracking, ERP, and beyond*. New York: Psychology Press.

Cedrus Corporation. (2017). Superlab 5. Retrieved from https://www.cedrus.com/superlab/

Centraal Bureau voor de Statistiek. (2018). *Jaarrapport Integratie 2018*. Retrieved from https://www.cbs.nl/nl-nl/publicatie/2018/47/jaarrapport-integratie-2018

Centraal Bureau voor de Statistiek. (2019). *Bevolking; leeftijd, migratieachtergrond, geslacht en regio, 1 januari*. Retrieved from CBS StatLine: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37713/table?ts=1568934836316

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658. https://doi.org/10.1121/1.1815131

Cohen, M. M., & Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech. In M. Magnenat-Thalmann & D. Thalmann (Eds.), *Models and Techniques in Computer Animation* (pp. 139–156). Tokyo: Springer-Verlag.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, *120*(5), 2421–2424. https://doi.org/10.1121/1.2229005

Cornips, L. (2002). Etnisch Nederlands in Lombok. In H. Bennis, G. Extra, P. Muysken, & J. Nortier (Eds.), *Een buurt in beweging: Talen en culturen in het Utrechtse Lombok en Transvaal* (pp. 285–302). Amsterdam: Stichting Beheer IISG.

Corretge, R. (2019). *Praat Vocal Toolkit*. Retrieved from http://www.praatvocaltoolkit.com

Cutler, A., & Norris, D. (2016). Bottoms up! How top-down pitfalls ensnare speech perception researchers, too. *Behavioral and Brain Sciences*, *39*, e236. https://doi.org/10.1017/S0140525X15002745

D'Onofrio, A. (2015). Persona-based information shapes linguistic perception: Valley Girls and California vowels. *Journal of Sociolinguistics*, *19*(2), 241–256. https://doi.org/10.1111/josl.12115

D'Onofrio, A. (2016). *Social meaning in linguistic perception* (PhD Dissertation). Retrieved from http://faculty.wcas.northwestern.edu/~akd2621/research.html

D'Onofrio, A. (2019). Complicating categories: Personae mediate racialized expectations of non-native speech. *Journal of Sociolinguistics*, *23*(4), 1–21.

De Meo, A. (2012). How credibile is a non-native speaker? Prosody and surroundings. In M. G. Busà & A. Stella (Eds.), *Methodological Perspectives on Second Language Prosody: Papers from ML2P 2012* (pp. 3–9). Padova: Libraria Editrice Università di Padova.

De Meo, A., Vitale, M., Pettorino, M., & Martin, P. (2011). Acoustic-perceptual credibility correlates of news reading by native and Chinese speakers of Italian. In W. S. Lee & E. Zee (Eds.), *The 17th International Congress of Phonetic Sciences* (pp. 1366–1369). Hong Kong: City University of Hong Kong.

DeJesus, J. M., Hwang, H. G., Dautel, J. B., & Kinzler, K. D. (2018). "American = English speaker" before "American = White": The development of children's reasoning about nationality. *Child Development*, *89*(5), 1752–1767. https://doi.org/10.1111/cdev.12845

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(1), 1–16. https://doi.org/10.1017/S0272263197001010

Devos, T., & Banaji, M. R. (2005). American = White? *Journal of Personality and Social Psychology*, *88*(3), 447–466. https://doi.org/10.1037/0022-3514.88.3.447

Dorleijn, M., & Nortier, J. (2006). Het Marokkaanse accent in het Nederlands: Marker of indicator? In T. Koole, J. Nortier, & B. Tahity (Eds.), *Artikelen van de Vijfde Sociolinguïstische Conferentie* (pp. 138–148). Delft: Eburon.

Drager, K. (2005). From bad to bed: The relationship between perceived age and vowel perception in New Zealand English. *Te Reo*, *48*, 55–68.

Drager, K. (2011). Speaker age and vowel perception. *Language and Speech*, *54*(1), 99–121. https://doi.org/10.1177/0023830910388017

Drager, K., & Kirtley, M. J. (2016). Awareness, salience, and stereotypes in exemplar-based models of speech production and perception. In A. M. Babel (Ed.), *Awareness and Control in Sociolinguistic Research* (pp. 1–24). Cambridge: Cambridge University Press.

Dragojevic, M., & Giles, H. (2016). I don't like you because you're hard to understand: The role of processing fluency in the language attitudes process. *Human Communication Research*, *42*(3), 396–420. https://doi.org/10.1111/hcre.12079

Eklund, I., & Traunmüller, H. (1997). Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. *Phonetica*, *54*(1), 1–21. https://doi.org/10.1159/000262207

Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, *40*(4), 481–492. https://doi.org/10.1044/jshd.4004.481

Faulkenberry, T. J. (2017). A single-boundary accumulator model of response times in an addition verification task. *Frontiers in Psychology*, *8*, 1225. https://doi.org/10.3389/fpsyg.2017.01225

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology*, *42*(1), 98–116. https://doi.org/10.1037/0012-1649.42.1.98

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology*, *17*(1), 74–147. https://doi.org/10.1080/10463280600681248

Fiedler, S., Keller, C., & Hanulíková, A. (2019). Social expectations and intelligibility of Arabic-accented speech in noise. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia* (pp. 3085–3089). Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, *39*, 1–77. https://doi.org/10.1017/S0140525X15000965

Floccia, C., Butler, J., Goslin, J., & Ellis, L. (2009). Regional and foreign accent processing in English: Can listeners adapt? *Journal of Psycholinguistic Research*, *38*(4), 379–412. https://doi.org/10.1007/s10936-008-9097-8

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1276–1293. https://doi.org/10.1037/0096-1523.32.5.1276

Foulkes, P. (2010). Exploring social-indexical knowledge: A long past but a short history. *Laboratory Phonology*, *1*(1), 5–39. https://doi.org/10.1515/labphon.2010.003

Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, *34*(4), 409–438. https://doi.org/10.1016/j.wocn.2005.08.002

Foulkes, P., & Hay, J. B. (2015). The emergence of sociophonetic structure. In B. MacWhinney & W. O'Grady (Eds.), *The Handbook of Language Emergence* (pp. 399–3100). https://doi.org/10.1002/9781118346136.ch13

Frances, C., Costa, A., & Baus, C. (2018). On the effects of regional accents on memory and credibility. *Acta Psychologica*, *186*, 63–70. https://doi.org/10.1016/j.actpsy.2018.04.003

Fry, E. (2011). *1000 Instant Words: The most common words for teaching reading, writing and spelling*. Westminster: Teacher Created Resources, Inc.

Fryer, R. G., & Levitt, S. D. (2004). The causes and consequences of distinctively black names. *The Quarterly Journal of Economics*, *119*(3), 767–805. https://doi.org/10.1162/0033553041502180

Fyock, J., & Stangor, C. (1994). The role of memory biases in stereotype maintenance. *British Journal of Social Psychology*, *33*(3), 331–343. https://doi.org/10.1111/j.2044-8309.1994.tb01029.x

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, *34*(1), 65–87. https://doi.org/10.1111/j.1467-1770.1984.tb00996.x

Glidden, C. M., & Assmann, P. F. (2004). Effects of visual gender and frequency shifts on vowel category judgments. *Acoustics Research Letters Online*, *5*(4), 132–138. https://doi.org/10.1121/1.1764472

Gluszek, A., & Dovidio, J. F. (2010). The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review*, *14*(2), 214–237. https://doi.org/10.1177/1088868309359288

Gnevsheva, K. (2018). The expectation mismatch effect in accentedness perception of Asian and Caucasian non-native speakers of English. *Linguistics*, *56*(3), 581–598. https://doi.org/10.1515/ling-2018-0006

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. https://doi.org/10.1037/0033-295X.105.2.251

Gonsalkorale, K., Sherman, J. W., & Klauer, K. C. (2009). Aging and prejudice: Diminished regulation of automatic race bias among older adults. *Journal of Experimental Social Psychology*, *45*(2), 410–414. https://doi.org/10.1016/j.jesp.2008.11.004

Gonzalez-Barrera, A., & Connor, P. (2019). Around the world, more say immigrants are a strength than a burden. Retrieved June 12, 2019, from Pew Research Center website: https://www.pewresearch.org/global/2019/03/14/around-the-world-more-say-immigrants-are-a-strength-than-a-burden/

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197–1208. https://doi.org/10.1121/1.422512

Grey, S., Schubel, L., McQueen, J., & van Hell, J. (2019). Processing foreign-accented speech in a second language: Evidence from ERPs during sentence comprehension in bilinguals. *Bilingualism: Language and Cognition*, *22*(5), 912–929. https://doi.org/10.1017/S1366728918000937

Grey, S., & van Hell, J. G. (2017). Foreign-accented speaker identity affects neural correlates of language comprehension. *Journal of Neurolinguistics*, *42*, 93–108. https://doi.org/10.1016/j.jneuroling.2016.12.001

Grondelaers, S., & Speelman, D. (2015). A quantitative analysis of qualitative free response data: Paradox or new paradigm? In J. Daems, E. Zenner, K. Heylen, D. Speelman, & H. Cuyckens (Eds.), *Change of Paradigms – New Paradoxes: Recontextualizing Language and Linguistics* (pp. 361–384). De Gruyter, Inc.

Grondelaers, S., & van Gent, P. (2019). How “deep” is Dynamism? Revisiting the evaluation of Moroccan-flavored Netherlandic Dutch. *Linguistics Vanguard*, *5*(s1), 1–11. https://doi.org/10.1515/lingvan-2018-0011

Grondelaers, S., van Gent, P., & van Hout, R. (2015). Is Moroccan-flavoured Standard Dutch standard or not? On the use of perceptual criteria to determine the limits of standard languages. In *Responses to Language Varieties: Variability, Processes and Outcomes* (pp. 191–218). https://doi.org/10.1075/impact.39.09gro

Guy, G. R. (2011). Sociolinguistics and formal linguistics. In R. Wodak, B. Johnstone, & P. Kerswill (Eds.), *The SAGE Handbook of Sociolinguistics* (pp. 249–264). London: SAGE Publications.

Hagendoorn, L., Veenman, J., & Vollebergh, W. (2003). Cultural orientation and socio-economic integration of immigrants in the Netherlands. In L. Hagendoorn, J. Veenman, & W. Vollebergh (Eds.), *Integrating Immigrants in the Netherlands* (pp. 1–16). Aldershot: Ashgate Publishing Limited.

Hansen, K., Rakić, T., & Steffens, M. C. (2017). Competent and warm?: How mismatching appearance and accent influence first impressions. *Experimental Psychology*, *64*(1), 27–36. https://doi.org/10.1027/1618-3169/a000348

Hansen, K., Rakić, T., & Steffens, M. C. (2018). Foreign-looking native-accented people: More competent when first seen rather than heard. *Social Psychological and Personality Science*, *9*(8), 1001–1009. https://doi.org/10.1177/1948550617732389

Hansen, K., Steffens, M. C., Rakić, T., & Wiese, H. (2017). When appearance does not match accent: neural correlates of ethnicity-related expectancy violations. *Social Cognitive and Affective Neuroscience*, *12*(3), 507–515. https://doi.org/10.1093/scan/nsw148

Hanulíková, A. (2018). The effect of perceived ethnicity on spoken text comprehension under clear and adverse listening conditions. *Linguistics Vanguard*, *4*(1), 1–9. https://doi.org/10.1515/lingvan-2017-0029

Hanulíková, A., van Alphen, P. M., van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, *24*(4), 878–88. https://doi.org/10.1162/jocn_a_00103

Hanzlíková, D., & Skarnitzl, R. (2017). Credibility of native and non-native speakers of English revisited: Do non-native listeners feel the same? *Research in Language*, *15*(3). https://doi.org/10.1515/rela-2017-0016

Harrison, G. (2014). Accent and "othering" in the workplace. In J. Levis & A. Moyer (Eds.), *Social Dynamics in Second Language Accent* (pp. 255–272). Boston: DeGruyter Mouton.

Hay, J., & Bresnan, J. (2006). Spoken syntax: The phonetics of giving a hand in New Zealand English. *Linguistic Review*, *23*(3), 321–349. https://doi.org/https://doi.org/10.1515/TLR.2006.013

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, *48*(4), 865–892. https://doi.org/10.1515/LING.2010.027

Hay, J., Nolan, A., & Drager, K. (2006). From fush to feesh: Exemplar priming in speech perception. *Linguistic Review*, *23*(3), 351–379. https://doi.org/10.1515/TLR.2006.014

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*(4), 458–484. https://doi.org/10.1016/j.wocn.2005.10.001

Hiebert, D. (2006). Winning, losing, and still playing the game: The political economy of immigration in Canada. *Tijdschrift Voor Economische En Sociale Geografie*, *97*(1), 38–48. https://doi.org/10.1111/j.1467-9663.2006.00494.x

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*(3), 253–269. https://doi.org/10.1080/01434630802651677

Ingvalson, E. M., Lansford, K. L., Fedorova, V., & Fernandez, G. (2017). Cognitive factors as predictors of accented speech perception for younger and older adults. *The Journal of the Acoustical Society of America*, *141*(6), 4652–4659. https://doi.org/10.1121/1.4986930

Jaspers, J. (2004). Marokkaanse jongens en het Antwerps dialect. *Taal En Tongval: Tijdschrift Voor de Taalvariatie*, *17*, 135–165.

Jaspers, J. (2005). Linguistic sabotage in a context of monolingualism and standardization. *Language & Communication*, *25*(3), 279–297. https://doi.org/10.1016/j.langcom.2005.03.007

Jensen, C., Denver, L., Mees, I. M., & Werther, C. (2013). Students' attitudes to lecturers' English in English-medium higher education in Denmark. *NJES Nordic Journal of English Studies*, *13*(1), 87–112.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, *34*(4), 485–499. https://doi.org/10.1016/j.wocn.2005.08.004

Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. In M.-J. Solé, P. S. Beddor, & M. Ohala (Eds.), *Experimental Approaches to Phonology* (pp. 25–40). Oxford: Oxford University Press.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384. https://doi.org/10.1006/jpho.1999.0100

Johnstone, B., & Kiesling, S. F. (2008). Indexicality and experience: Exploring the meanings of /aw/-monophthongization in Pittsburgh. *Journal of Sociolinguistics*, *12*(1), 5–33. https://doi.org/10.1111/j.1467-9841.2008.00351.x

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

Kang, O., & Rubin, D. L. (2014). Listener expectations, reverse linguistic stereotyping, and individual background factors in social judgments and oral performance assessment. In J. M. Levis & A. Moyer (Eds.), *Social dynamics in second language accent* (pp. 239–254). Boston: Walter de Gruyter.

Karpinska, M. (2019). How accented do Caucasian-looking vs. Asian-looking native speakers sound to a Japanese listener? In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia* (pp. 3691–3695). Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Kaup, B., Lüdtke, J., & Zwaan, R. (2005). Effects of negation, truth value, and delay on picture recognition after reading affirmative and negative sentences. *Proceedings of the Annual Meeting of the Cognitive Science Society, 27*. Retrieved from https://escholarship.org/uc/item/19s068vb

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, *64*(3), 459–489. https://doi.org/10.3138/cmlr.64.3.459

Kim, J. (2016). Perceptual associations between words and speaker age. *Laboratory Phonology: Journal of the Association of Laboratory Phonology*, *7*(1), 1–22. https://doi.org/10.5334/labphon.33

Kirby, J., & Sonderegger, M. (2018). Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics*, *70*, 70–85. https://doi.org/10.1016/j.wocn.2018.05.005

Kist, R. (2015a, July 27). Jouw steenkolen-Engels kan een deal verprutsen. *NRCQ*. Retrieved from https://www.nrc.nl/nieuws/2015/07/27/jouw-steenkolen-engels-kan-een-deal-verprutsen-a1495589

Kist, R. (2015b, July 28). Houterig accent? 1-0 achter. *NRC*. Retrieved from https://www.nrc.nl/nieuws/2015/07/28/houterig-accent-1-0-achter-1521645-a647817

Kleinschmidt, D. F., Weatherholtz, K., & Jaeger, T. F. (2018). Sociolinguistic perception as inference under uncertainty. *Topics in Cognitive Science*, *10*(4), 818–834. https://doi.org/10.1111/tops.12331

Klok, P. (2008). “Je gaat je kapot lachen. Woela.” *De Volkskrant*. Retrieved from https://www.volkskrant.nl/nieuws-achtergrond/je-gaat-je-kapot-lachen-woela~b58f7926/?referer=https%3A%2F%2Fwww.google.com%2F

Koops, C., Gentry, E., & Pantos, A. (2008). The effect of perceived speaker age on the perception of PIN and PEN vowels in Houston, Texas. *University of Pennsylvania Working Papers in Linguistics*, *14*(2), 91–101. Retrieved from http://repository.upenn.edu/pwpl/vol14/iss2/12

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Kymlicka, W. (2012). *Multiculturalism: Success, failure, and the future*. Washington, D.C.: Migration Policy Institute.

Labov, W. (2001). Applying our knowledge of African American English to the problem of raising reading levels in inner-city schools. In S. L. Lanehart (Ed.), *Sociocultural and Historical Contexts of African American English* (pp. 299–318). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Lambert, W. E., Anisfeld, M., & Yeni-Komshian, G. (1965). Evaluation reactions of Jewish and Arab adolescents to dialect and language variations. *Journal of Personality and Social Psychology*, *2*(1), 84–90. https://doi.org/10.1037/h0022088

Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *Journal of Abnormal and Social Psychology*, *60*(1), 44–51. https://doi.org/10.1037/h0044430

Lattner, S., & Friederici, A. D. (2003). Talker's voice and gender stereotype in human auditory sentence processing: Evidence from event-related brain potentials. *Neuroscience Letters*, *339*(3), 191–194. https://doi.org/10.1016/S0304-3940(03)00027-2

Lenth, R. V. (2019). *Estimated Marginal Means, aka Least-Squares Means [R package]*. Retrieved from https://github.com/rvlenth/emmeans

Lev-Ari, S., Ho, E., & Keysar, B. (2018). The unforeseen consequences of interacting with non-native speakers. *Topics in Cognitive Science*, *10*(4), 835–849. https://doi.org/10.1111/tops.12325

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, *46*(6), 1093–1096. https://doi.org/10.1016/j.jesp.2010.05.025

Lev-Ari, S., & Peperkamp, S. (2016). How the demographic makeup of our community influences speech perception. *The Journal of the Acoustical Society of America*, *139*(6), 3076–3087. https://doi.org/10.1121/1.4950811

Levi, S. V., Winters, S. J., & Pisoni, D. B. (2007). Speaker-independent factors affecting the perception of foreign accent in a second language. *The Journal of the Acoustical Society of America*, *121*(4), 2327–2338. https://doi.org/10.1121/1.2537345

Lindemann, S. (2002). Listening with an attitude: A model of native-speaker comprehension of non-native speakers in the United States. *Language in Society*, *31*(3), 419–441. https://doi.org/10.1017/S0047404502020286

Lippi-Green, R. (2012). *English with an accent: language, ideology, and discrimination in the United States* (2nd ed.). London; New York: Routledge.

Little, W. (2016). Race and ethnicity. In *Introduction to Sociology - 2nd Canadian edition*. Vancouver: BC Campus Open Source Textbook. Retrieved from https://opentextbc.ca/introductiontosociology2ndedition/chapter/chapter-11-race-and-ethnicity/.

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition*, *174*, 55–70. https://doi.org/10.1016/j.cognition.2018.01.003

Maas, E., & Mailend, M.-L. (2012). Speech planning happens before speech execution: Online reaction time methods in the study of apraxia of speech. *Journal of Speech, Language, and Hearing Research*, *55*(5), 1523–1534. https://doi.org/10.1044/1092-4388(2012/11-0311)

MacWhinney, B., Feldman, H., Sacco, K., & Valdes-Perez, R. (2000). Online measures of basic language skills in children with early focal brain lesions. *Brain and Language*, *71*(3), 400–431. https://doi.org/10.1006/brln.1999.2273 [doi]\nS0093-934X(99)92273-3 [pii]

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391–409. https://doi.org/10.1037/0096-1523.33.2.391

Mahtani, M., & Mountz, A. (2002). *Immigration to British Columbia: Media representation and public opinion* (No. 02–15). Retrieved from http://mbc.metropolis.net/assets/uploads/files/wp/2002/WP02-15.pdf

Marquer, J., & Pereira, M. (1990). Reaction times in the study of strategies in sentence–picture verification: A reconsideration. *The Quarterly Journal of Experimental Psychology Section A*, *42*(1), 147–168. https://doi.org/10.1080/14640749008401212

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

May, L., Baron, A. S., & Werker, J. F. (2019). Who can speak that language? Eleven-month-old infants have language-dependent expectations regarding speaker ethnicity. *Developmental Psychobiology*, *61*(6), 859–873. https://doi.org/10.1002/dev.21851

McCrocklin, S., Blanquera, K., & Loera, D. (2018). Student perceptions of university instructor accent in a linguistically diverse area. In J Levis (Ed.), *Proceedings of the 9th Pronunciation in Second Language Learning and Teaching conference, Salt Lake City, USA* (pp. 141–150). Ames, IA: Iowa State University.

McGowan, K. B. (2011). *The role of socioindexical expectation in speech perception (Doctoral dissertation)*. Retrieved from https://deepblue.lib.umich.edu/handle/2027.42/89684

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech*, *58*(4), 502–521. https://doi.org/10.1177/0023830914565191

McGowan, K. B. (2016). Sounding Chinese and listening Chinese: Awareness and knowledge in the laboratory. In A. M. Babel (Ed.), *Awareness and Control in Sociolinguistic Research* (pp. 25–61). Cambridge: Cambridge University Press.

McGowan, K. B., & Babel, A. M. (2019). Perceiving isn't believing: Divergence in levels of sociolinguistic awareness. *Language in Society*, 1–26. https://doi.org/10.1017/S0047404519000782

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

McLennan, C. T., & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 306–321. https://doi.org/10.1037/0278-7393.31.2.306

Miller, C. L., Younger, B. A., & Morse, P. A. (1982). The categorization of male and female voices in infancy. *Infant Behavior and Development*, *5*(2–4), 143–159. https://doi.org/10.1016/S0163-6383(82)80024-6

Millotte, S., René, A., Wales, R., & Christophe, A. (2008). Phonological phrase boundaries constrain the online syntactic analysis of spoken sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 874–885. https://doi.org/10.1037/0278-7393.34.4.874

Munro, M. J. (2003). A primer on accent discrimination in the Canadian context. *TESL Canada Journal*, *20*(2), 38–51. https://doi.org/10.18806/tesl.v20i2.947

Munro, M. J., & Derwing, T. M. (1995). Processing Time, Accent, and Comprehensibility in the Perception of Native and Foreign-Accented Speech. *Language and Speech*, *38*(3), 289–306. https://doi.org/10.1177/002383099503800305

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility Of L2 speech. *Studies in Second Language Acquisition*, *28*(1), 111–131. https://doi.org/10.1017/S0272263106060049

Munson, B., Jefferson, S. V., & McDonald, E. C. (2006). The influence of perceived sexual orientation on fricative identification. *The Journal of the Acoustical Society of America*, *119*(4), 2427–2437. https://doi.org/10.1121/1.2173521

Münster, K., & Knoeferle, P. (2018). Extending situated language comprehension (accounts) with speaker and comprehender characteristics: Toward socially situated interpretation. *Frontiers in Psychology*, *8*(2267), 1–12. https://doi.org/10.3389/fpsyg.2017.02267

Neuhauser, S., & Simpson, A. P. (2007). Imitated or authentic? Listeners' judgements of foreign accents. *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany*, 1805–1808. Retrieved from http://www.icphs2007.de

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*(1), 62–85. https://doi.org/10.1177/0261927X99018001005

Nortier, J. (2017). Gender-related online metalinguistic comments on Straattaal and Moroccan Flavored Dutch in the Moroccan heritage community in the Netherlands. *Applied Linguistics Review*, *10*, 341–366. https://doi.org/10.1515/applirev-2017-0047

Nortier, J., & Dorleijn, M. (2008). A Moroccan accent in Dutch: A sociocultural style restricted to the Moroccan community? *International Journal of Bilingualism*, *12*(1–2), 125–142. https://doi.org/10.1177/13670069080120010801

Nortier, J., & Dorleijn, M. (2017). "Het is een grappige accent weet je." Retrieved October 2, 2019, from NEMO kennislink website: https://www.nemokennislink.nl/publicaties/het-is-een-grappige-accent-weet-je/

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Pantos, A. (2019). Implicitness, automaticity, and consciousness in language attitudes research. *Linguistics Vanguard*, *5*(s1), 1–9. https://doi.org/10.1515/lingvan-2018-0007

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006

Perception Research Systems. (2007). *Paradigm Stimulus Presentation*. Retrieved from http://www.paradigmexperiments.com

Perry, L. K., Mech, E. N., MacDonald, M. C., & Seidenberg, M. S. (2018). Influences of speech familiarity on immediate perception and final comprehension. *Psychonomic Bulletin & Review*, *25*(1), 431–439. https://doi.org/10.3758/s13423-017-1297-5

Pharao, N., & Kristiansen, T. (2019). Reflections on the relation between direct/indirect methods and explicit/implicit attitudes. *Linguistics Vanguard*, *5*(s1), 1–7. https://doi.org/10.1515/lingvan-2018-0010

R Core Team. (2014). *R: A language and environment for statistical computing*. Retrieved from http://www.r-project.org/

Rácz, P. (2013). *Salience in sociolinguistics: A quantitative approach*. Berlin: De Gruyter Mouton.

Riach, P. A., & Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal*, *112*(483), F480–F518. https://doi.org/10.1111/1468-0297.00080

Romero-Rivas, C., Martin, C. D., & Costa, A. (2015). Processing changes when listening to foreign-accented speech. *Frontiers in Human Neuroscience*, *9*(167), 1–15. https://doi.org/10.3389/fnhum.2015.00167

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, *33*(4), 511–531. https://doi.org/10.1007/BF00973770

Rubin, D. L. (2012). The power of prejudice in accent perception: Reverse linguistic stereotyping and its impact on listener judgements and decisions. In John Levis & K. LeVelle (Eds.), *Proceedings of the 3rd Annual Pronunciation in Second Language Learning and Teaching Conference* (pp. 11–17). https://doi.org/10.13140/RG.2.1.1465.4485

Rubin, D. L., Ainsworth, S., Cho, E., Turk, D., & Winn, L. (1999). Are Greek letter social organizations a factor in undergraduates' perceptions of international instructors? *International Journal of Intercultural Relations*, *23*(1), 1–12. https://doi.org/10.1016/S0147-1767(98)00023-6

Rubin, D. L., Coles, V. B., & Barnett, J. T. (2016). Linguistic stereotyping in older adults' perceptions of health care aides. *Health Communication*, *31*(7), 911–916. https://doi.org/10.1080/10410236.2015.1007549

Rubin, D. L., & Smith, K. A. (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. https://doi.org/10.1016/0147-1767(90)90019-S

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., & Simola, J. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, *127*(1), 141–145. https://doi.org/10.1016/0304-3940(91)90914-F

Schmid, M. (2019). Here's how your foreign accent can unfairly destroy your credibility. Retrieved from the Conversation website: https://theconversation.com/heres-how-your-foreign-accent-can-unfairly-destroy-your-credibility-125981

Schmid, P. M., & Yeni-Komshian, G. H. (1999). The effects of speaker accent and target predictability on perception of mispronunciations. *Journal of Speech Language and Hearing Research*, *42*(1), 56–64. https://doi.org/10.1044/jslhr.4201.56

Senior, B., Hui, J., & Babel, M. (2018). Liu vs. Liu vs. Luke? Name influence on voice recall. *Applied Psycholinguistics*, *39*(6), 1117–1146. https://doi.org/10.1017/S0142716418000267

Shapiro, L., Swinney, D., & Borsky, S. (1998). Online examination of language performance in normal and neurologically impaired adults. *American Journal of Speech-Language Pathology*, *7*(1), 49–60. https://doi.org/10.1044/1058-0360.0701.49

Singal, J. (2017). Psychology's favorite tool for measuring racism isn't up to the job. https://doi.org/10.1111/sjop.12288

Smith, E. R., & Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, *99*(1), 3–21. https://doi.org/10.1037/0033-295X.99.1.3

Souza, A. L., & Markman, A. B. (2013). Foreign accent does not influence cognitive judgments. *Proceedings of the Annual Meeting of the Cognitive Science Society, 35*, 1360–1365. Retrieved from https://escholarship.org/uc/item/8hd9v4ff

Squires, L. (2013). It don't go both ways: Limited bidirectionality in sociolinguistic perception. *Journal of Sociolinguistics*, *17*(2), 200–237. https://doi.org/10.1111/josl.12025

Statistics Canada. (2017a). *Greater Vancouver, RD [Census division], British Columbia and British Columbia [Province] (table). Census Profile.* Statistics Canada Catalogue no. 98-316-X2016001. Ottawa. Released November 29, 2017. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E (accessed December 2, 2019).

Statistics Canada. (2017b). *Greater Vancouver, RD [Census division], British Columbia and British Columbia [Province] (table). Census Profile*. 2016 Census. Statistics Canada Catalogue no. 98-316-X2016001. Ottawa. Released November 29, 2017. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E (accessed September 16, 2019).

Statistics Canada. (2017c). *Immigration and Ethnocultural Diversity Highlight Tables*. 2016 Census. Statistics Canada Catalogue no 98-402-X2016007. Ottawa. Released February 14, 2018. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/imm/Table.cfm?Lang=E&T=42&Geo=59 (accessed September 16, 2019).

Staum Casasanto, L. (2008). Does social information influence sentence processing? In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society, Washington* (pp. 799–804). Austin, TX: Cognitive Science Society.

Staum Casasanto, L. (2010). What do listeners know about sociolinguistic variation? *University of Pennsylvania Working Papers in Linguistics*, *15*(2), 40–49.

Staum Casasanto, L., Grondelaers, S., & van Hout, R. (2015). Got class? Community-shared conceptualizations of social class in evaluative reactions to sociolinguistic variables. In A. Prikhodkine & D. R. Preston (Eds.), *Responses to language varieties: Variability, processes and outcomes* (pp. 159–173). https://doi.org/10.1075/impact.39.07cas

Steenkolenengels: not so very believable. (2015, July 28). *RTL Nieuws*. Retrieved from https://www.rtlnieuws.nl/editienl/artikel/1052716/steenkolenengels-not-so-very-believable

Stewart, B. D., von Hippel, W., & Radvansky, G. A. (2009). Age, race, and implicit prejudice. *Psychological Science*, *20*(2), 164–168. https://doi.org/10.1111/j.1467-9280.2009.02274.x

Stocker, L. (2017). The impact of foreign accent on credibility: An analysis of cognitive statement ratings in a Swiss context. *Journal of Psycholinguistic Research*, *46*(3), 617–628. https://doi.org/10.1007/s10936-016-9455-x

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology*, *18*(1), 86–100. https://doi.org/10.1177/0261927X99018001006

Strand, E. A., & Johnson, K. (1996). Gradient and Visual Speaker Normalization in the Perception of Fricatives. In D. Gibbon (Ed.), *Natural Language Speech Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielefeld, October, 1996* (pp. 318–336). https://doi.org/10.1515/9783110821895-003

Sumner, M., & Kataoka, R. (2013). Effects of phonetically-cued talker variation on semantic encoding. *The Journal of the Acoustical Society of America*, *134*(6), EL485–EL491. https://doi.org/10.1121/1.4826151

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology*, *4*(JAN), 1–13. https://doi.org/10.3389/fpsyg.2013.01015

Tekin, O. (2019). The association between ethnic group affiliation and the ratings of comprehensibility, intelligibility, accentedness, and acceptability. *The Electronic Journal of English as a Second Language*, *23*(3).

The Asia Pacific Foundation of Canada. (2018). *National opinion poll 2018: Canadian views on Asia*. Retrieved from https://www.asiapacific.ca/sites/default/files/filefield/nop2018_0.pdf

Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning*, *41*(2), 177–204. https://doi.org/10.1111/j.1467-1770.1991.tb00683.x

Uttley, L., de Boisferon, A. H., Dupierrix, E., Lee, K., Quinn, P. C., Slater, A. M., & Pascalis, O. (2013). Six-month-old infants match other-race faces with a non-native language. *International Journal of Behavioral Development*, *37*(2), 84–89. https://doi.org/10.1177/0165025412467583

van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, *20*(4), 580–591. https://doi.org/10.1162/jocn.2008.20054

van Oudenhoven, J. P., Prins, K. S., & Buunk, B. P. (1998). Attitudes of minority and majority members towards adaptation of immigrants. *European Journal of Social Psychology*, *28*(6), 995–1013. https://doi.org/10.1002/(SICI)1099-0992(1998110)28:6<995::AID-EJSP908>3.0.CO;2-8

van Oudenhoven, J. P., Ward, C., & Masgoret, A. M. (2006). Patterns of relations between immigrants and host societies. *International Journal of Intercultural Relations*. https://doi.org/10.1016/j.ijintrel.2006.09.001

Vaughn, C., & Baese-Berk, M. (2019). Effects of talker order on accent ratings. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia* (pp. 1253–1257). Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Von Hippel, W., Silver, L. A., & Lynch, M. E. (2000). Stereotyping against your will: The role of inhibitory ability in stereotyping and prejudice among the elderly. *Personality and Social Psychology Bulletin*, *26*(5), 523–532. https://doi.org/10.1177/0146167200267001

Walker, A., & Hay, J. (2011). Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology*, *2*(1), 219–237. https://doi.org/10.1515/labphon.2011.007

Walker, M., Szakay, A., & Cox, F. (2019). Can kiwis and koalas as cultural primes induce perceptual bias in Australian English speaking listeners? *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *10*(1), 1–29. https://doi.org/10.5334/labphon.90

Weatherhead, D., & White, K. S. (2018). And then I saw her race: Race-based expectations affect infants' word processing. *Cognition*, *177*, 87–97. https://doi.org/10.1016/j.cognition.2018.04.004

Weil, S. A. (2001). *Foreign accented speech: Encoding and generalization (Master's thesis)*. Retrieved from https://www.researchgate.net/publication/265924633_Foreign_Accented_Speech_Encoding_and_Generalization

Weil, S. A. (2003). The impact of phonetic dissimilarity on the perception of foreign accented speech. *The Journal of the Acoustical Society of America*, *114*(4), 2423–2423. https://doi.org/10.1121/1.4778795

White, M. (2018). How big should the control group be in a randomized field experiment? [blog post]. In *Mark H. White II, PhD*. Retrieved from https://www.markhw.com/blog/control-size

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

Wondershare. (2019). Wondershare Filmora 9. Retrieved from https://filmora.wondershare.com/

Xie, Z., Yi, H.-G., & Chandrasekaran, B. (2014). Nonnative audiovisual speech perception in noise: Dissociable effects of the speaker and listener. *PLoS ONE*, *9*(12), 1–17. https://doi.org/10.1371/journal.pone.0114439

Yi, H.-G., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2013). Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of the Acoustical Society of America*, *134*(5), EL387–EL393. https://doi.org/10.1121/1.4822320

Yi, H.-G., Smiljanic, R., & Chandrasekaran, B. (2014). The neural processing of foreign-accented speech and its relationship to listener bias. *Frontiers in Human Neuroscience*, *8*, 1–12. https://doi.org/10.3389/fnhum.2014.00768

Yook, C., & Lindemann, S. (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. https://doi.org/10.1080/01434632.2012.734509

Zahn, C. J., & Hopper, R. (1985). Measuring language attitudes: The speech evaluation instrument. *Journal of Language and Social Psychology*, *4*(2), 113–123. https://doi.org/10.1177/0261927X8500400203

Zangl, R., & Fernald, A. (2007). Increasing flexibility in children's online processing of grammatical and nonce determiners in fluent speech. *Language Learning and Development*, *3*(3), 199–231. https://doi.org/10.1080/15475440701360564

Zhao, B., Ondrich, J., & Yinger, J. (2006). Why do real estate brokers continue to discriminate? Evidence from the 2000 Housing Discrimination Study. *Journal of Urban Economics*, *59*(3), 394–419. https://doi.org/10.1016/j.jue.2005.12.001

Zheng, Y., & Samuel, A. G. (2017). Does seeing an Asian face make speech sound more accented? *Attention, Perception, and Psychophysics*, *79*(6), 1841–1859. https://doi.org/10.3758/s13414-017-1329-2

Zheng, Y., & Samuel, A. G. (2019). How much do visual cues help listeners in perceiving accented speech? *Applied Psycholinguistics*, *40*(1), 93–109. https://doi.org/10.1017/S0142716418000462

Appendix A.

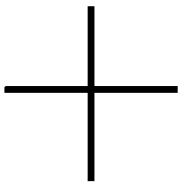
True/False sentences used in Experiment 1

1. Hot and cold are opposites.
2. Elephants are big animals.
3. Exercise is good for your health.
4. Shakespeare wrote many fine plays.
5. Some people love to eat chocolate.
6. Some people keep dogs as pets.
7. Young children can be very noisy.
8. Hungry cats like to chase mice.
9. You can start a fire with a match.
10. Red and green are colors.
11. Many houses are made of bricks.
12. France is a country in Europe.
13. Gold is a valuable metal / material.
14. Ships travel on the water.
15. You can buy a burger at McDonalds.
16. A city is bigger than a village.
17. Computers can be very useful.
18. Italy is known for its good food.
19. A team works together.
20. One is a smaller number than two.
21. A baby depends on its parents.
22. Language is one way to communicate.
23. A face can show expression.
24. Yesterday is in the past.
25. Christmas is in December.
26. The word 'immediately' means now.
27. You need money to buy things.
28. The weather can change suddenly.
29. Gasoline is an excellent drink.
30. The Queen of England lives in Washington.

31. Fish live in tall trees.
32. The inside of an egg is blue.
33. August is a winter month.
34. It always snows in July.
35. Most people wear hats on their feet.
36. The stars come out in the day.
37. Wednesday is the first day of the week.
38. All men can have babies.
39. All dogs have fifteen legs.
40. People eat through their noses /nose.
41. A quarter is worth ten cents.
42. There are many cities on the moon.
43. You can buy beer at church.
44. A monkey is a kind of bird.
45. March has thirty-four days.
46. Eating breakfast is dangerous.
47. Modern and old are the same.
48. Parents are younger than their children.
49. Friends don't care about you.
50. Everyone is looking forward to death.
51. There are six minutes in an hour.
52. Arms are much longer than legs.
53. Something cheap is always of high quality.
54. When you feel fine you go to the hospital.
55. The future has already happened.
56. Everything always goes as expected.

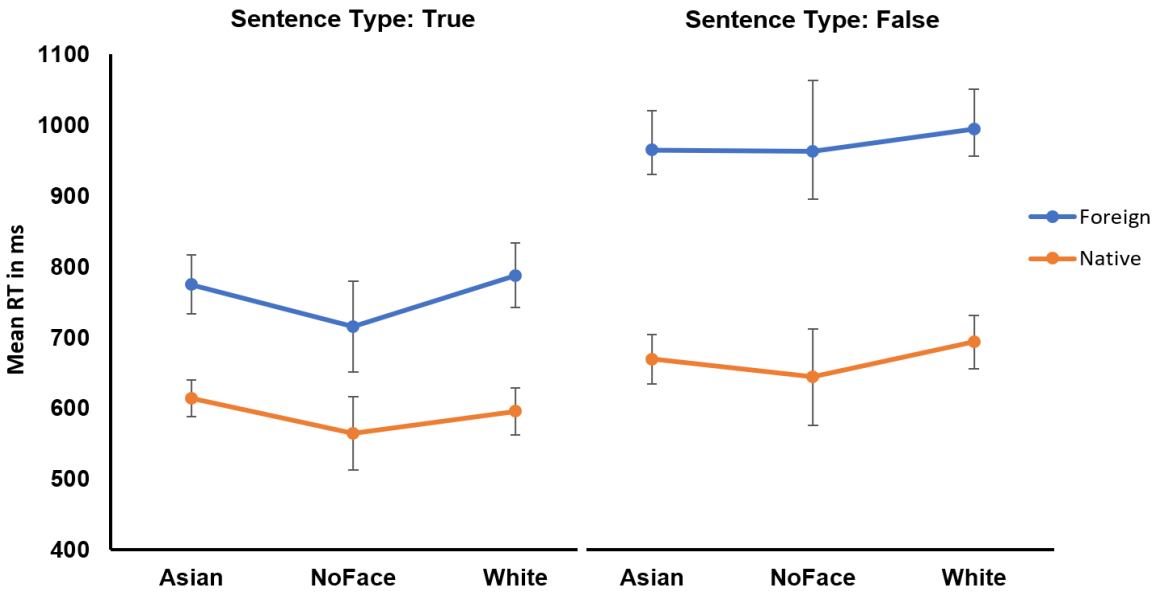
Appendix B.

Photos used in Experiment 1



Appendix C.

Mean RTs with SE bars for all levels of Face and Veracity



Appendix D.

True/False sentences used in Experiment 2

1. Hot and cold are opposites.
2. It is healthy to exercise.
3. Shakespeare wrote many plays.
4. Some people love to eat chocolate.
5. Young children can be noisy.
6. Hungry cats like to chase mice.
7. You can start a fire with a match.
8. Red and green are colors.
9. A ring can be made out of gold.
10. Ships travel on the water.
11. A city is bigger than a village.
12. Computers can be very useful.
13. Italy is known for its good food.
14. A team works together.
15. A face can show expression.
16. Yesterday is in the past.
17. Christmas is in December.
18. You need money to buy things.
19. You can sit on a chair.
20. You can get money from a bank.
21. A forest has a lot of trees.
22. Ice is made of water.
23. The earth is a planet.
24. A race car can go very fast.
25. Students go to school.
26. You can make music with an instrument.
27. Scientists often run experiments.
28. You can swim in a lake.
29. Gasoline is an excellent drink.
30. The Queen of England lives in Washington.

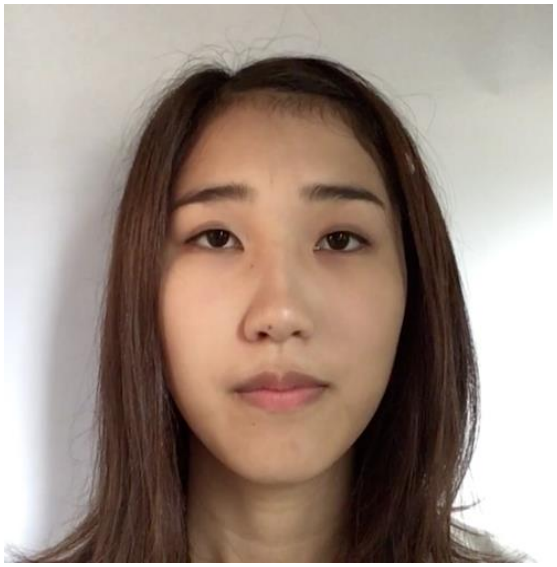
31. People wear hats on their feet.
32. People eat through their noses.
33. There are many cities on the moon.
34. You can buy paint at church.
35. A monkey is a kind of bird.
36. There are seven days in March.
37. Eating breakfast is dangerous.
38. Parents are younger than their children.
39. Friends don't care about you.
40. Everyone likes to eat paper.
41. There are six minutes in an hour.
42. A pineapple is a vegetable.
43. When you feel fine you go to the hospital.
44. Everything always goes as expected.
45. You can walk on water.
46. The shape of a football is square.
47. A river is larger than the ocean.
48. You can fit a horse in your shoe.
49. Mountains are very small.
|50. Dictionaries describe history.
51. You leave your house through the window.
52. You can learn a language in a day.
53. You wear warm clothes in summer.
54. A baby knows how to read.
55. Vancouver is a country.
56. Fish have thick hair.

Practice sentences:
Most apples are blue.
Cows live in tall trees.
Black bears live in forests.
A baby depends on its parents.

Appendix E.

Video-recorded speakers for Experiment 2



Appendix F.

Estimates of fixed effect regression coefficients for the linear mixed effects model with corresponding *p*-values and the variable coding scheme (Experiment 2)

	Estimate	SE	df	*t*	Pr(>|t|)
(Intercept)	6.803	0.061	346	112.304	0.000
Face_NoFace	-0.205	0.095	214	-2.162	0.032
Face_White	-0.031	0.069	318	-0.454	0.650
Voice ID_Foreign2	0.050	0.053	371	0.953	0.341
Voice ID_Native1	-0.077	0.050	306	-1.544	0.124
Voice ID_Native2	-0.067	0.054	311	-1.242	0.215
Veracity_True : Trial	-0.199	0.057	194	-3.459	0.001
Trial	0.012	0.001	334	11.384	0.000
Face_NoFace : Voice ID_Foreign2	0.035	0.088	360	0.399	0.690
Face_White : Voice ID_Foreign2	0.074	0.079	367	0.943	0.346
Face_NoFace : Voice ID_Native1	-0.091	0.084	309	-1.082	0.280
Face_White : Voice ID_Native1	-0.026	0.077	312	-0.338	0.736
Face_NoFace : Voice ID_Native2	-0.060	0.087	313	-0.694	0.488
Face_White : Voice ID_Native2	0.017	0.084	305	0.207	0.836
Face_NoFace : Veracity_True	0.061	0.070	221	0.873	0.383
Face_White : Veracity_True	0.089	0.056	210	1.589	0.114
Face_NoFace : Trial	-0.013	0.001	708	-10.341	0.000
Face_White : Trial	0.002	0.001	270	1.070	0.286
Voice ID_Foreign2 : Veracity_True	0.038	0.060	256	0.639	0.523
Voice ID_Native1 : Veracity_True	0.062	0.054	187	1.149	0.252
Voice ID_Native2 : Veracity_True	0.119	0.055	195	2.172	0.031
Veracity_True : Trial	0.002	0.001	353	2.210	0.028
Face_NoFace : Voice ID_Foreign2 : Veracity_True	0.040	0.107	276	0.374	0.709
Face_White : Voice ID_Foreign2 : Veracity_True	-0.052	0.084	256	-0.612	0.541
Face_NoFace : Voice ID_Native1 : Veracity_True	-0.058	0.096	204	-0.601	0.548
Face_White : Voice ID_Native1 : Veracity_True	-0.077	0.078	194	-0.989	0.324
Face_NoFace : Voice ID_Native2 : Veracity_True	-0.264	0.097	207	-2.713	0.007
Face_White : Voice ID_Native2 : Veracity_True	-0.173	0.078	196	-2.231	0.027

Random effects

Groups	Variance	Std.Dev.
PFVA (intercept)	0.002	0.043
PFV (intercept)	0.009	0.093
Participant (intercept)	0.040	0.200
Sentence.Nr (intercept)	0.018	0.134
Residual (intercept)	0.106	0.325

Variable coding scheme

$X_1 = 1$	if *Face* is NoFace	$X_1 = 0$	if *Face* is Moroccan or White
$X_2 = 1$	if *Face* is White	$X_2 = 0$	if *Face* is White or NoFace
$X_3 = 1$	if *Voice* is Foreign2	$X_3 = 0$	if *Voice* is not Foreign2
$X_4 = 1$	if *Voice* is Native1	$X_4 = 0$	if *Voice* is not Native1
$X_5 = 1$	if *Voice* is Native2	$X_5 = 0$	if *Voice* is not Native2
$X_6 = 1$	Trial number * estimated regression coefficient for *Trial*		

Appendix G.

RTs for each *Face–Voice ID* at *Trials* 15 and 43

Estimated median RTs for each *Face–Voice ID* combination at *Trial* = 15

					95% Confidence Interval		
Voice ID	Face	Median	SE	df	Lower	Upper	Group*
Native2	Audio	556	39.02	117	484.09	639.14	1
Native1	Audio	575	40.11	115	500.33	659.72	12
Foreign1	Audio	679	47.77	118	590.42	780.22	23
Foreign2	Audio	769	55.92	135	665.78	887.75	34
Native1	White	912	42.32	208	832.76	999.85	45
Native1	Asian	939	43.42	207	857.29	1028.72	45
Native2	White	945	43.4	205	862.7	1034.08	45
Native2	Asian	977	45.88	216	890.49	1071.62	45
Foreign1	Asian	984	46.33	216	896.55	1079.46	456
Foreign1	White	1020	48.32	220	928.58	1119.32	56
Foreign2	Asian	1055	50.96	239	958.78	1159.88	56
Foreign2	White	1147	54.91	234	1043.97	1260.65	6

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Estimated median RTs for each *Face–Voice ID* combination at *Trial* = 43

					95% Confidence Interval		
Voice ID	Face	Median	SE	df	Lower	Upper	Group*
Native2	Audio	547	38.3	115.9	476.19	628.39	1
Native1	Audio	565	39.67	117.26	491.65	649.3	12
Foreign1	Audio	667	47.19	120.44	580.29	767.76	23
Foreign2	Audio	756	54.42	129.13	655.71	871.76	3
Native1	Asian	1333	62.24	210.75	1216.07	1461.78	4
Native1	White	1351	63.59	213.7	1231.45	1482.5	4
Native2	Asian	1387	65.79	216.79	1263.1	1522.8	4
Foreign1	Asian	1397	65.09	213.79	1274.09	1531.07	4
Native2	White	1399	65.27	210.22	1275.64	1533.34	4
Foreign2	Asian	1497	73.23	245.85	1359.67	1648.58	45
Foreign1	White	1510	71.02	217.52	1375.93	1656.28	45
Foreign2	White	1699	80.57	230.66	1547.15	1865.12	5

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Appendix H.

Probability tables at *Trials* 29 and 43

Probability that one *Face–Voice ID* combination takes longer than the other at *Trial* = 29

					95% Confidence Interval		Probability interval in %	
Voice	Comparison	Ratio	SE	df	Lower	Upper	Lower	Upper
Foreign1	Asian vs. White	0.945	0.044	191	0.811	1.101	43.5	52.5
Foreign2	Asian vs. White	0.900	0.043	220	0.767	1.056	40.9	50.9
Native1	Asian vs. White	1.008	0.046	174	0.867	1.171	46.0	54.5
Native2	Asian vs. White	1.013	0.046	182	0.870	1.178	46.1	54.8

Probability that the AV condition takes longer than the AO condition at *Trial* = 29

					95% Confidence Interval		Probability interval in %	
Voice	Comparison	Ratio	SE	df	Lower	Upper	Lower	Upper
Foreign1	Asian vs. Audio	1.742	0.137	113	1.339	2.265	63.1	72.9
Foreign1	White vs. Audio	1.843	0.043	114	1.415	2.401	74.8	65.1
Foreign2	Asian vs. Audio	1.648	0.134	128	1.257	2.161	62.5	73.8
Foreign2	White vs. Audio	1.831	0.044	126	1.398	2.398	76.8	66.1
Native1	Asian vs. Audio	1.964	0.153	109	1.513	2.549	66.4	75.5
Native1	White vs. Audio	1.949	0.040	110	1.500	2.532	75.6	66.4
Native2	Asian vs. Audio	2.110	0.166	111	1.623	2.743	68.8	77.8
Native2	White vs. Audio	2.084	0.038	110	1.604	2.706	77.2	68.2

Probability that one *Face–Voice ID* combination takes longer than the other at *Trial* = 43

					95% Confidence Interval		Probability interval in %	
Voice	Comparison	Ratio	SE	df	Lower	Upper	Lower	Upper
Foreign1	Asian vs. White	0.925	0.046	196	0.785	1.090	42.9	51.8
Foreign2	Asian vs. White	0.881	0.046	227	0.742	1.047	40.3	50.3
Native1	Asian vs. White	0.987	0.049	188	0.836	1.165	45.3	53.8
Native2	Asian vs. White	0.992	0.05	195	0.838	1.173	45.4	54.0

Probability that the AV condition takes longer than the AO condition at *Trial* = 43

					95% Confidence Interval		Probability interval in %	
Voice	Comparison	Ratio	SE	df	Lower	Upper	Lower	Upper
Foreign1	Asian vs. Audio	2.092	0.169	123	1.599	2.739	68.1	77.4
Foreign1	White vs. Audio	2.262	0.036	124	1.727	2.963	70.7	79.7
Foreign2	Asian vs. Audio	1.980	0.165	137	1.502	2.611	68.1	78.7
Foreign2	White vs. Audio	2.247	0.037	133	1.708	2.955	72.1	81.9
Native1	Asian vs. Audio	2.360	0.190	121	1.805	3.085	70.8	79.4
Native1	White vs. Audio	2.391	0.034	122	1.828	3.129	71.4	80.0
Native2	Asian vs. Audio	2.535	0.205	121	1.938	3.318	73.1	81.5
Native2	White vs. Audio	2.557	0.031	119	1.957	3.341	73.1	81.4

Appendix I.

True/False sentences used in Experiment 3

1. Een koelkast houdt eten langer vers. / A refrigerator keeps food fresh for longer.
2. Een plant heeft licht nodig. / A plant needs light.
3. Computers zijn erg nuttig. / Computers are very useful.
4. Italië staat bekend om het goede eten. / Italy is known for its good food.
5. Gisteren is in het verleden. / Yesterday is in the past.
6. Een olifant heeft een slurf. / An elephant has a trunk.
7. Je kan dingen kopen met geld. / You can buy things with money.
8. Stamppot is typisch Nederlands. / Stamppot is typically Dutch.
9. Je kan geld opnemen bij de bank. / You can withdraw money at the bank.
10. Er staan veel bomen in een bos. / There are many trees in a forest.
11. Water dat bevriest wordt ijs. / Water that freezes becomes ice.
12. De aarde is een planet. / The earth is a planet.
13. Er wordt gezongen in musicals. / There is singing in musicals.
14. Leerlingen gaan naar school. / Students go to school.
15. Een eiland is omringd door water. / An island is surrounded by water.
16. Een trein rijdt op het spoor. / A train runs on the tracks.
17. Je kan varen op een meer. / You can sail on a lake.
18. Het is gezond om te sporten. / It is healthy to exercise.
19. Sommige mensen houden van chocolade. / Some people love chocolate.
20. Een hond is een huisdier. / A dog is a pet.
21. We kijken met onze ogen. / We look with our eyes.
22. Rood and groen zijn kleuren. / Red and green are colors.
23. Een ring kan gemaakt zijn van goud. / A ring can be made of gold.
24. Een baby is afhankelijk van zijn ouders. / A baby depends on its parents.
25. De mens heeft tien vingers. / Humans have ten fingers.
26. Een ganzenveer is zwaar. / A goose feather is heavy.
27. Een aap is een soort vogel. / A monkey is a kind of bird.
28. Iedereen eet graag papier. / Everybody likes to eat paper.
29. Er zitten zes maanden in een jaar. / There are six months in a year.
30. Een banaan is een groente. / A banana is a vegetable.

31. Gezonde mensen gaan naar het ziekenhuis. / Healthy people go to the hospital.
32. Een kat heeft vleugels. / A cat has wings.
33. Sinaasappelsap is gemaakt van katoen. / Orange juice is made from cotton.
34. Een voetbal is vierkant. / A football is square.
35. Een zandkasteel bouw je van sneeuw. / You can build a sandcastle from snow.
36. Lekkere koekjes smaken naar plastic. / Tasty cookies taste like plastic.
37. De meeste appels zijn blauw. / Most apples are blue.
38. Een miljonair is arm. / A millionaire is poor.
39. Je verlaat je huis via het raam. / You leave your house through the window.
40. Je kan een taal leren binnen een minuut. / You can learn a language within a minute.
41. We dragen warme kleren in de zomer. / We wear warm clothes in the summer.
42. Een baby kan zelf lezen. / A baby knows how to read.
43. Amsterdam is een land. / Amsterdam is a country.
44. Glas kan niet breken. / Glass cannot break.
45. Vissen hebben lang haar. / Fish have long hair.
46. Ouders zijn jonger dan hun kinderen. / Parents are younger than their children.
47. Kokend water is koud. / Boiling water is cold.
48. De fiets is sneller dan het vliegtuig. / Bikes are faster than planes.
49. Er zijn veel steden op de maan. / There are many cities on the moon.
50. Ontbijten is gevaarlijk. / Having breakfast is dangerous.

Appendix J.

Trivia statements used in Experiment 3

1. Kamelen hebben drie oogleden. / Camels have three eyelids.
2. Koeien kunnen niet zweten. / Cows cannot sweat.
3. Mieren hebben geen longen. / Ants don't have lungs.
4. Emoes zijn de enige vogelsoorten met kuitspieren. / Emus are the only bird species with calf muscles.
5. Tijgers hebben een gestreepte huid. / Tigers have striped skin.
6. Het hart van een garnaal bevindt zich in zijn kop. / A shrimp's heart is located on his head.
7. Kikkers kunnen niet overgeven. / Frogs cannot vomit.
8. De Amerikaanse vlag is ontworpen door een 17-jarige. / The American flag was designed by a 17-year-old.
9. Een giraf hoeft minder vaak te drinken dan een kameel. / A giraffe needs to drink less often than a camel
10. De eerste fiets had geen remmen. / The first bike had no brakes.
11. Rusland heeft een groter landoppervlak dan Pluto. / Russia has a larger surface area than Pluto.
12. De VS is een ouder land dan Duitsland. / The US is an older country than Germany.
13. De elektrische stoel is uitgevonden door een tandarts. / The electric chair was invented by a dentist.
14. Het regent diamanten op Jupiter en Saturnus. / It rains diamonds on Jupiter and Saturn.
15. De blikopener werd pas 48 jaar na het blikje uitgevonden. / The can opener was only invented 48 years after the can.
16. De meest gebruikte kleur in landenvlaggen is rood. / The most used color in national flags is red.
17. Een mug heeft 47 tanden. / A mosquito has 47 teeth.
18. Mensen met veel moedervlekken hebben meer kans op huidkanker. / People with many birthmarks are more likely to develop skin cancer.
19. Er zitten ongeveer 800 druiven in een fles wijn. / There are approximately 800 grapes in a bottle of wine.
20. China is de grootste knoflookproducent ter wereld. / China is the largest producer of garlic in the world.
21. Honing bederft niet. / Honey does not spoil.

22. Het kanon werd uitgevonden door een Franse monnik. / The cannon was invented by a French monk.
23. De keizerspinguïn kan maximaal tien minuten onder water blijven. / The emperor penguin can stay under water for a maximum of ten minutes.
24. Een kat heeft 23 spieren in elk oor. / A cat has 23 muscles in each ear.
25. Valken paren iedere keer met een ander vrouwtje. / Falcons mate with a different female every time.
26. Een kakkerlak kan bijna twee maanden zonder voedsel leven. / A cockroach can live without food for almost two months.
27. Volwassen insecten zijn kleurenblind. / Adult insects are color blind.
28. Een slak kan tot twee maanden slapen. / A snail can sleep for up to two months.
29. Leonardo da vinci heeft de schaar uitgevonden. / Leonardo da Vinci invented scissors.
30. Steve Jobs heeft de muis uitgevonden. / Steve Jobs invented the computer mouse.
31. Ierland heeft de meeste brouwerijen ter wereld. / Ireland has the most breweries in the world.
32. Vleermuizen zijn volledig blind. / Bats are completely blind.
33. Vlinders sterven als je hun vleugels aanraakt. / Butterflies die when you touch their wings.
34. De kameleon verandert van kleur om zichzelf te kamoufleren. / The chameleon changes color to camouflage itself.
35. Schildpadden voelen geen pijn door hun schild. / Turtles can feel no pain through their shield.
36. Koeien kunnen de trap niet aflopen. / Cows cannot go down the stairs.
37. Geluid verplaatst zich langzamer door water dan door de lucht. / Sound travels more slowly through water than through air.
38. Diamanten kunnen niet verbranden. / Diamonds cannot burn.
39. Het eerste land dat postzegels verkocht was Duitsland. / The first country to sell stamps was Germany.
40. De eerste webcam was gericht op een winkelstraat. / The first webcam showed a shopping street.
41. Regenwater is goed om te drinken. / Rainwater is safe to drink.
42. Smaakpapillen worden elke maand vernieuwd. / Taste buds are renewed every month.
43. De kleinste botten in het menselijk lichaam bevinden zich in de hand. / The smallest bones in the human body are in the hand.

Appendix K.

Video-recorded speakers for Experiment 3












Appendix L.

Estimates of fixed effect regression coefficients for the linear mixed effects model with corresponding *p*-values and the variable coding scheme (Experiment 3)

	Estimate	SE	df	*t*	Pr(>|t|)
(Intercept)	5.967	0.120	171	49.933	0.000
Face_White	0.224	0.094	172	2.377	0.019
Face_Moroccan	0.194	0.095	177	2.037	0.043
Voice_Native2	0.201	0.085	743	2.381	0.018
Voice_Foreign1	0.550	0.094	976	5.852	0.000
Voice_Foreign2	0.417	0.086	779	4.854	0.000
Veracity_True	-0.140	0.092	219	-1.511	0.132
Age	0.008	0.002	112	3.394	0.001
Face_White:Voice_Native2	-0.109	0.078	3187	-1.387	0.165
Face_Moroccan:Voice_Native2	-0.020	0.077	3186	-0.254	0.800
Face_White:Voice_Foreign1	-0.149	0.088	3249	-1.688	0.092
Face_Moroccan:Voice_Foreign1	-0.045	0.089	3264	-0.510	0.610
Face_White:Voice_Foreign2	-0.081	0.078	3213	-1.034	0.301
Face_Moroccan:Voice_Foreign2	-0.057	0.079	3215	-0.719	0.472
Face_White:Veracity_True	-0.001	0.079	803	-0.008	0.994
Face_Moroccan:Veracity_True	0.018	0.081	887	0.228	0.820
Voice_Native2:Veracity_True	0.050	0.103	435	0.488	0.625
Voice_Foreign1:Veracity_True	-0.014	0.113	570	-0.122	0.903
Voice_Foreign2:Veracity_True	0.083	0.103	438	0.799	0.425
Voice_Native2:Age	-0.003	0.001	3270	-2.353	0.019
Voice_Foreign1:Age	-0.002	0.002	3204	-1.564	0.118
Voice_Foreign2:Age	-0.005	0.001	3261	-3.449	0.001
Face_White:Voice_Native2:Veracity_True	0.001	0.109	3177	0.012	0.991
Face_Moroccan:Voice_Native2:Veracity_True	-0.080	0.108	3171	-0.747	0.455
Face_White:Voice_Foreign1:Veracity_True	0.068	0.121	3230	0.558	0.577
Face_Moroccan:Voice_Foreign1:Veracity_True	-0.108	0.123	3247	-0.878	0.380
Face_White:Voice_Foreign2:Veracity_True	-0.061	0.108	3204	-0.564	0.573
Face_Moroccan:Voice_Foreign2:Veracity_True	-0.070	0.110	3202	-0.634	0.526

Random effects

Groups	Variance	Std.Dev.
SFV (intercept)	0.000	0.000
SV (intercept)	0.023	0.152
SF (intercept)	0.003	0.053
Participant (intercept)	0.097	0.311
Sentence Nr (intercept)	0.037	0.191
Residual	0.208	0.456

Variable coding scheme

$X_1 = 1$	if *Face* is White	$X_3 = 0$	if *Face* is Moroccan or NoFace
$X_2 = 1$	if *Face* is Moroccan	$X_2 = 0$	if *Face* is White or NoFace
$X_3 = 1$	if *Voice* is Native2	$X_3 = 0$	if *Voice* is not Native2
$X_4 = 1$	if *Voice* is Foreign1	$X_4 = 0$	if *Voice* is not Foreign1
$X_5 = 1$	if *Voice* is Foreign2	$X_5 = 0$	if *Voice* is not Foreign2
$X_6 = 1$	if *Veracity* is True	$X_6 = 0$	if *Veracity* is False
$X_7 =$	Age of participants * estimated regression coefficient for *Age*		

Appendix M.

Accentedness ratings for each *Face–Voice* pairing at *Ages* 23 and 31

Model-estimated mean accentedness ratings for each *Face–Voice ID* combination at *Age* = 23

					95% Confidence Interval		
Face	Voice ID	Mean	SE	df	Lower	Upper	Group*
White	Native2	1.020	0.222	324	0.582	1.46	1
Audio	Native1	1.120	0.219	320	0.690	1.55	1
Audio	Native2	1.136	0.219	320	0.706	1.57	1
Moroccan	Native2	1.225	0.197	324	0.837	1.61	1
White	Native1	1.311	0.227	324	0.865	1.76	1
Moroccan	Native1	1.478	0.193	324	1.098	1.86	1
Moroccan	Foreign2	5.643	0.231	324	5.189	6.10	2
Audio	Foreign2	5.848	0.220	310	5.415	6.28	2
Audio	Foreign1	6.295	0.219	320	5.865	6.73	2
White	Foreign2	6.302	0.191	324	5.925	6.68	2
White	Foreign1	6.373	0.199	324	5.982	6.76	2
Moroccan	Foreign1	6.564	0.219	324	6.134	6.99	2

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Model-estimated mean accentedness ratings for each *Face–Voice ID* combination at *Age* = 31

					95% Confidence Interval		
Face	Voice ID	Mean	SE	df	Lower	Upper	Group*
White	Native2	1.055	0.174	324	0.713	1.40	1
Audio	Native1	1.093	0.183	320	0.733	1.45	1
Audio	Native2	1.096	0.183	320	0.736	1.46	1
Moroccan	Native2	1.208	0.160	324	0.894	1.52	1
White	Native1	1.255	0.176	324	0.908	1.60	1
Moroccan	Native1	1.361	0.162	324	1.043	1.68	1
Moroccan	Foreign2	5.349	0.178	324	4.999	5.70	2
Audio	Foreign2	6.113	0.184	313	5.750	6.48	23
White	Foreign2	6.197	0.158	324	5.886	6.51	3
Audio	Foreign1	6.303	0.183	320	5.942	6.66	3
White	Foreign1	6.336	0.162	324	6.018	6.65	3
Moroccan	Foreign1	6.449	0.172	324	6.112	6.79	3

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Appendix N.

Voice ID effects at *Ages* 30 and 52

Median response times in ms at each level of *Voice ID* for *Age* = 30

				95% Confidence Interval		
Voice ID	Mean RT	SE	df	Lower	Upper	Group*
Native1	533	28.1	181	480	592	1
Native2	573	30.1	180	516	635	1
Foreign2	680	35.9	182	613	755	2
Foreign1	793	43.1	202	712	883	3

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Median response times in ms at each level of *Voice ID* for *Age* = 52

				95% Confidence Interval		
Voice ID	Mean RT	SE	df	Lower	Upper	Group*
Native2	635	40.3	177	560	720	1
Native1	636	40.4	179	561	720	1
Foreign2	728	46.3	179	642	825	2
Foreign1	895	59.8	217	784	1021	3

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Appendix O.

Voice ID and *Face* effects at *Ages* 30 and 52

Model-estimated median response times for each *Face–Voice* pairing at *Age* = 30

					95% Confidence Interval		
Voice ID	Face	Median RT	SE	df	Lower	Upper	Group*
Native1	Audio	462	37.0	156	395	541	1
Native2	Audio	525	41.8	154	449	615	12
Native1	Moroccan	566	34.9	230	502	640	1 3
Native1	Dutch	579	35.8	233	512	654	1 3
Native2	Dutch	590	36.6	237	522	667	1 3
Native2	Moroccan	606	36.9	219	538	683	1234
Foreign2	Audio	631	50.3	155	539	739	345
Foreign2	Moroccan	706	43.9	238	624	798	2 45
Foreign2	Dutch	707	43.5	228	626	798	2 45
Foreign1	Audio	739	61.4	182	627	870	345
Foreign1	Moroccan	819	53.0	275	721	931	5
Foreign1	Dutch	824	53.0	267	726	935	5

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Model-estimated median response times for each *Face–Voice* pairing at *Age* = 52

					95% Confidence Interval		
Voice ID	Face	Median RT	SE	df	Lower	Upper	Group*
Native1	Audio	551	50.4	146	460	660	1
Native2	Audio	582	53.1	145	486	697	12
Native2	Dutch	654	45.6	228	570	750	123
Native2	Moroccan	672	45.1	198	589	767	123
Native1	Moroccan	675	47.4	234	588	775	123
Foreign2	Audio	676	61.6	145	564	809	234
Native1	Dutch	690	46.4	199	604	788	123
Foreign2	Moroccan	755	51.6	212	660	864	123
Foreign2	Dutch	756	52.4	223	660	867	123
Foreign1	Audio	833	80.2	178	689	1008	34
Foreign1	Moroccan	924	66.6	258	802	1065	4
Foreign1	Dutch	930	67.9	272	805	1074	4

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Appendix P.

Credibility scores for *Voice ID* and *Face*

Model-estimated median credibility score at each level of *Voice*

				95% Confidence Interval		
Voice ID	Median	SE	df	Lower	Upper	Group*
Native2	3.722	0.153	58	3.417	4.027	1
Foreign1	3.754	0.152	58	3.449	4.059	1
Native1	3.771	0.152	58	3.466	4.076	1
Foreign2	3.870	0.152	58	3.564	4.175	1

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Model-estimated median credibility score at each level of *Face*

				95% Confidence Interval		
Face	Median	SE	df	Lower	Upper	Group*
Audio-Only	3.731	0.154	58	3.424	4.038	1
Moroccan	3.774	0.146	50	3.480	4.068	1
White	3.832	0.147	50	3.538	4.127	1

* Conditions that share the same number in the 'Group' column are not significantly different from one another.

Model-estimated median credibility score at each combination of *Face* and *Voice*

					95% Confidence Interval		
Face	Voice ID	Median	SE	df	Lower	Upper	Group*
Audio	Native2	3.661	0.179	107	3.306	4.016	1
Moroccan	Native2	3.697	0.167	82	3.366	4.029	1
Audio	Foreign1	3.721	0.179	107	3.365	4.076	1
Moroccan	Native1	3.735	0.171	90	3.396	4.074	1
Audio	Foreign2	3.748	0.179	107	3.393	4.103	1
White	Foreign1	3.770	0.168	84	3.436	4.104	1
Moroccan	Foreign1	3.771	0.170	88	3.434	4.108	1
White	Native1	3.783	0.168	84	3.450	4.117	1
Audio	Native1	3.794	0.179	108	3.438	4.149	1
White	Native2	3.808	0.173	94	3.465	4.151	1
Moroccan	Foreign2	3.892	0.170	89	3.554	4.230	1
White	Foreign2	3.969	0.169	86	3.633	4.305	1

* Conditions that share the same number in the 'Group' column are not significantly different from one another.