# Do L2 Speakers' Assumptions About the Test Examiner Influence Their Speaking Anxiety in an Oral Exam? A Reverse Linguistic Stereotyping Study

Lin Lu

A Thesis
In the Department of
Education

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Arts (Applied Linguistics) at
Concordia University
Montreal, Quebec, Canada

July 2025

**CONCORDIA UNIVERSITY**

**SCHOOL OF GRADUATE STUDIES**

This is to certify that the thesis prepared

By: Lin Lu

Entitled: Do L2 Speakers' Assumptions About the Test Examiner Influence Their Speaking Anxiety in an Oral Exam? A Reverse Linguistic Stereotyping Study

and submitted in partial fulfillment of the requirements for the degree of

**Master of Arts (Applied Linguistics)**

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair

Dr. Rachael Lindberg Examiner

Dr. Pavel Trofimovich Thesis Supervisor

Approved by

Dr. Walcir Cardoso Graduate Program Director

Dr. Pascale Sicotte Dean of Faculty

## Abstract

Do L2 Speakers' Assumptions About the Test Examiner Influence Their Speaking Anxiety in an Oral Exam? A Reverse Linguistic Stereotyping Study

Lin Lu

A speech may be perceived differently depending on what social information we draw from the speaker, be it correct or misguided (Burgers & Beukeboom, 2020; Edwards, 1999; Lambert et al., 1960; Niedzielski, 1999). The phenomenon where non-linguistic social information about the speaker (e.g., race, occupation, etc.) influences our actual experience with speech is called reverse linguistic stereotyping (RLS; Kang & Rubin, 2009). Although RLS has been well documented for its influence on speech perception (e.g., seeing an Asian face can render a speech less comprehensible and more accented), its broader impact on speakers is still less known. This study, therefore, examined how L2 speakers' assumptions about examiners influence their speaking anxiety in an oral exam.

Participants included 40 Mandarin-speaking international students in Montreal, who completed two English speaking tests delivered through video prompts. Each test featured a different examiner (Caucasian or South Asian), while the audio remained constant (Canadian English). Participants rated their speaking anxiety before and after each test and evaluated each examiner. Retrospective recall interviews were conducted with eight individuals who showed noticeable difference in pretest anxiety ratings across two examiners. Results revealed no significant difference in speaking anxiety in Test 1. However, a higher pretest anxiety was observed in Test 2 when the examiner appeared South Asian than Caucasian, possibly due to a shift in visual stimuli, which activated the stereotypical association between examiner's race and

linguistic ability. The findings highlighted the importance of creating a more inclusive environment in both language learning and assessment.

**Acknowledgements**

The completion of this thesis marks one of the happiest and most fulfilling moments in the past two years, and I am deeply grateful to everyone who helped me along the way. I would first like to express my deepest gratitude to my supervisor, Prof. Pavel Trofimovich, who supported me generously in every possible way — from the visualization of this work to detailed and timely feedback on every draft, to painstaking help with data analysis, and much more. Most importantly, I want to thank Pavel for leading by example the kind of researcher I hope to become.

I also feel extremely lucky and grateful to have had Prof. Rachael Lindberg as my committee member — this thesis benefited greatly from her insights and encouragement, and reading her work on speaking anxiety was truly inspiring for me.

Thank you as well to the professors and friends I met in the APLI program — for the exchange of wonderful ideas in class, for believing in my work, for studying together, and for the laughter and companionship we shared outside the classroom. Thank you to my Concordia Applied Linguistics Lab for being a safe place where I could learn, share, and grow. I would also like to extend my sincere gratitude to every participant in my study for your time, interest, and willingness to share your experiences with me.

Lastly, I want to say a big thank you to my partner and my family. Words cannot do justice to how much I love you and appreciate your love and support. I am the luckiest person in the world to have you in my life. And this last thank you goes to my sister, who, from the day we were born, will always share equally in every achievement of mine.

## Table of Contents

## List of Figures

## List of Tables

## Introduction

When it comes to teaching English as a second language (L2), there is often a preference for teachers who are first language (L1) speakers (Baratta, 2017; Clark & Paran, 2007), especially in many Asian countries (e.g., Chun, 2014; Galloway, 2014). For example, among 59 websites that recruit English teachers for China, Japan, Korea, Taiwan, and Thailand, 81% of the ads included a requirement for the chosen teacher to be an L1 speaker of English (Ruecker & Ives, 2015). Additionally, many L2 learners prefer L1 speakers as teachers, especially when it comes to developing L2 speaking skills, and learners often wish to speak like an L1 speaker of English (Lasagabaster & Sierra, 2002; Levis et al., 2016). Indeed, many people firmly believe that there exists an idealized English speaker whose English is more accurate, standard, and legitimate than that of speakers of other English varieties (Higgins, 2003; Lippi-Green, 2011).

The notion of L1 speaker as an ideal language model is a complex phenomenon (Cook, 1999; Rampton, 1990). The L1 versus L2 dichotomy is frequently driven by people's perception of language, for instance, in terms of speakers' accent or their use of "standard" grammar and vocabulary. However, a growing body of literature suggests that the widespread preference for L1 speakers also reflects an association of language with race, where L1 speakers are often conceptualized as individuals who are predominantly White and who reside in affluent western industrialized nations (Golombek & Jordan, 2005; Kubota & Lin, 2006; Rampton, 1990). For instance, English learners from various linguistic backgrounds in Canada mostly preferred a White teacher because they believed that only White people are L1 English speakers and only they possess the "real" Canadian English (Amin, 1997). In Japan, non-White English teachers (including a Black woman from South Africa, a Korean woman, and a Filipino woman) reported experience of marginalization in their English teaching career (Simon-Maeda, 2004). In fact, the

association between race and language appears so deep-rooted that seeing an Asian face might create an expectation for listeners to "hear" an L2 accent, even if no accented speech is actually involved (Babel & Russell, 2015; Gnevsheva, 2018; Kutlu et al., 2022; Rubin, 1992). This phenomenon, known as reverse linguistic stereotyping (RLS), reflects a perceptual bias in which listeners' assumptions about a speaker's linguistic ability are influenced not by the speech itself, but by visual or social cues associated with the speaker's identity (Kang & Rubin, 2009).

The stereotypical association between racial identity and language skills has been widely observed, and several consequences of this association for people's comprehension and their reactions to language have been documented (Babel & Russell, 2015; Gnevsheva, 2018; Hu & Lindemann, 2009; Hu & Su, 2015; Kang & Rubin, 2009; Kutlu et al., 2022; Rubin, 1992). Nevertheless, various other potential consequences of this association still require research, especially for L2 speakers. One such consequence concerns L2 speakers' emotional reactions—and more specifically, their L2 speaking anxiety—when interacting with individuals who might be either preferred or dispreferred as L1 speakers. The goal of this study is, therefore, to capture L2 speakers' emotional reactions (i.e., speaking anxiety) when completing a speaking assessment task administered by individuals from different racial backgrounds (Caucasian vs. South Asian).

## Literature Review

### Social Evaluation and Speech Perception

People hold different attitudes toward different languages. Although people's preference for one language over another might be based on the specific characteristics of a given language such as its aesthetic or functional value as perceived by each person (Giles et al., 1979; Hilton et al., 2022), language attitudes frequently reflect the accumulated social and affective evaluations that people have about different speech communities (Burgers & Beukeboom, 2020; Giles et al., 1974; Lambert et al., 1960; Tekin & Trofimovich, 2023). In one classic study, Lambert et al. (1960) used a matched-guise technique to elicit language attitudes toward English and French speakers in Quebec. Not knowing that they were listening to bilingual speakers sometimes using English and sometimes French, residents of Quebec evaluated less favourably the speaker they heard in the French than in the English guise. These reactions likely captured listeners' negative attitudes toward Quebec French speakers, most of whom belonged to the working class at the time. In a more recent study, Schüppert et al. (2015) reported that a Swedish–Danish bilingual speaker was rated more favourably by the Danes when she spoke Swedish than by the Swedes when she spoke Danish. The asymmetrical language attitudes toward Swedish and Danish possibly reflected a historically more influential status of Sweden as a political and economic power (Hilton et al., 2022).

Unsurprisingly, similar language attitudes also extend to speakers of different varieties of the same language (Ryan, 1983). For example, in a study of Danish students' attitudes toward speakers of British, American, and Australian English, Ladegaard (1998) found that the British speaker speaking the standard variety (Received Pronunciation) was rated most favourably in terms of both social status and various personal characteristics. Such attitudes are presumably the

product of a learned association between a given speech variety and the power possessed by those who have historically used it, including highly educated, wealthy, upper-class members of the British society.

Although early studies targeted attitudinal differences across the Inner Circle Englishes such as British, American, and Canadian varieties (e.g., Abrams & Hogg, 1987; Campbell-Kibler, 2005), recent research has targeted various Outer Circle Englishes, including those from India, Japan, and Hong Kong (e.g., Hansen et al., 2018; McKenzie, 2008). For example, Dragojevic and Goatley-Soan (2022) found that not only did American listeners attribute highest status and personality evaluations to Standard American English speakers compared to nine other Outer Circle English speakers, but they also perceived speakers of some varieties (e.g., French, German) more favourably than speakers of other varieties (e.g., Arabic, Farsi, Vietnamese), suggesting a nuanced social categorization of "foreignness" (Lee & Fiske, 2006; Lippi-Green, 2011). In fact, a preference for speakers of the Inner Circle English varieties appears to be widespread, for example, considering that Japanese listeners also rate speakers of American and British English more favourably than speakers of Japanese English in terms of status and competence (McKenzie, 2008). These findings reflect the traditional view of Inner Circle English varieties as superior to local varieties, which essentially mirrors people's stereotypical views about the social and ethnic groups speaking these language varieties (Hansen et al., 2018; Ryan, 1983).

Just as listeners use a speaker's speech to attribute social judgments to the speaker, they also draw on other cues to create various assumptions about the speaker and their speech performance. In an early study of this phenomenon, Detroit residents who listened to exactly the same speaker reported hearing more examples of Canadian raising (e.g., a stereotypically

Canadian way of pronouncing words like *about*, *ice*, and *house*) when they were led to believe that the speaker was Canadian, and they perceived fewer cases of Canadian raising when they believed that the speaker was a fellow Detroiter (Niedzielski, 1999). In essence, information about the speaker's geographic origin appeared to distort listeners' speech perception. In another striking demonstration, Hay and Drager (2010) found that listeners' perception of vowels could change simply as a result of which social category was activated by presenting listeners with a stuffed toy such as a kangaroo (a symbol for Australia) versus a kiwi (a symbol for New Zealand).

The phenomenon where non-linguistic social information about the speaker influences listeners' actual experience with speech is called reverse linguistic stereotyping (RLS). In essence, listeners stereotype "in reverse," not from speech to attitudes but rather from assumptions and preconceived ideas back to speech (for review, see Kang & Rubin, 2009). In a recent study, for instance, Kang et al. (2023) found that listeners perceived the same accented speech differently depending on the presumed occupation of the speaker (i.e., doctor, teacher, or waiter). Specifically, when listeners believed that the speaker was a university teacher, they rated the speech as less acceptable and comprehensible and considered the speaker as more accented, reflecting their high, and probably overly harsh, expectations for the teaching profession.

RLS effects are particularly salient in response to a speaker's visual cues to racial identity. In a classic study, Rubin (1992) employed the matched-guise technique to present a short academic lecture to L1 English university students. The lecture audios, recorded by an L1 English speaker, were paired with either an image of a Caucasian or an Asian woman, leading listeners to believe that the instructor had different racial identities. When listeners assumed that the instructor was Asian, they reported hearing a stronger accent and labeled the speech as less

standard. More strikingly, listeners also showed lower comprehension of the lecture presented in the Asian guise, suggesting that their actual experience with speech was impaired by their assumptions. Similar findings were observed by Hu and Su (2015) in their listening comprehension study with L2 learners. The learners who were told that the listening test was recorded by an American speaker outperformed those who were led to believe that the recording came from a Cantonese speaker of English. More recently, Kutlu et al. (2022) matched speech samples recorded by American, British, and Indian English speakers with images of White or South Asian people. Listeners transcribed the speech from all English varieties less accurately in the presence of a South Asian face than a White face, implying that speech intelligibility was impaired for listeners by their assumptions of a speaker's ethnicity (see also Kennedy et al., 2024). What was more intriguing is that the intelligibility of Indian English increased when the audio was paired with a White face, leading the researchers to conclude that "Whiteness as a construct governs listeners' engagement in speech" (p. 22).

**Second Language Anxiety**

Thus far, RLS research has focused on people's evaluation of speakers' linguistic performance, for example, in terms of the accentedness or intelligibility of their speech (Hay & Drager; 2010; Hu & Lindemann, 2009; Hu & Su, 2015; Kennedy et al., 2024; Kutlu et al., 2022; Niedzielski, 1999), or people's reactions to speakers along personal and professional dimensions such as competence, friendliness, and credibility (Gnevsheva, 2018; Kang et al., 2023; Kang & Rubin, 2009; Lee & Bailey, 2023). However, less well understood is how RLS impacts people, for instance, in terms of their affective or emotional states. One such affective state which might be susceptible to RLS effects is a speaker's anxiety, which refers to the "subjective, consciously perceived feelings of apprehension and tension, accompanied by or associated with activation or

arousal of the autonomic nervous system" (Spielberger, 1966, p. 17). It is important to distinguish between anxiety as a transitory state and a relatively stable personality trait. Whereas trait anxiety characterizes individuals who tend to experience anxiety more acutely as part of most daily activities (Spielberger, 1966; Spielberger et al., 1971), state anxiety happens in the moment, meaning that it is often a temporary reaction or response to specific and often unpleasant situations (Gregersen et al., 2014; Scovel, 1978; Spielberger et al., 1971).

When it comes to research on L2 anxiety, some scholars have attempted to identify anxious L2 speakers by looking at their unique background characteristics (e.g., Dewaele et al., 2008; Woodrow, 2006). However, most L2 research has focused on state anxiety, for instance, by identifying and describing anxiety-provoking situations such as speaking in a language classroom or in front of large groups of people (e.g., Hashemi, 2011; Liu, 2006). In fact, Horwitz et al. (1986) claimed that learning and communicating in an L2 is anxiety-provoking in and of itself, mainly because it "challenges an individual's self-concept as a competent communicator and leads to reticence, self-consciousness, fear, or even panic" (p. 128). Similarly, according to Cohen and Norst (1989), learning a new language is fundamentally different from learning other skills or subjects because of the "language fear" it engenders (p. 62). For these reasons, language anxiety is considered a separate form of situation-specific anxiety that should be studied in its own right (Horwitz, 2010).

Among different situations that specifically trigger language anxiety, the most cited one is when L2 speakers feel they are being tested or evaluated (Horwitz et al., 1986). In the classroom, for example, L2 speakers generally feel anxious when asked to speak in front of the whole class (e.g., giving a presentation), because they worry about negative evaluations from peers and teachers or believe that their language skills are not sufficiently advanced (Liu, 2006;

Mak, 2011). Indeed, L2 speakers experience greater anxiety and more intense feelings of failure when a learning activity is presented in the form of a test, especially when the test leads to a grade (Horwitz et al., 1986). In the same vein, Young (1986) pointed out that L2 learners' anxiety was not as high as expected if they knew that testing had no negative consequences for them. In addition, L2 anxiety emerges in an assessment context because of time pressure. For example, participants in Bielak's (2022) study reported greater anxiety when they needed to audio-record their answer within a 2-minute time limit compared to when they participated in a free discussion. According to Horwitz et al. (1986), because L2 communication is anxiety-inducing in itself, an oral test tends to evoke two subcomponents of situation-specific language anxiety simultaneously—test anxiety and oral communication anxiety.

Another situation-specific variable with consequences for L2 speakers' anxiety concerns the identity of their interlocutor, considering that anxiety could also be triggered in response to a specific person. When studying immigrant women's language experience in Canada, Peirce (1995), for instance, pointed out unequal power relations between immigrant women and speakers of the majority language, where the women considered themselves illegitimate speakers of English and experienced a sense of inferiority, coupled with feelings of anxiety, when interacting with anglophone Canadians. Enhanced levels of anxiety have been reported for L2 speakers when they communicate with individuals who they believe have the full command on the language, including L1 speakers (Woodrow, 2006) and language teachers (Hashemi, 2011; Horwitz et al., 1986; Mak, 2011). Speakers' perception of interlocutors' status and familiarity also plays a role in triggering speaking anxiety. For instance, Shirvan and Talebzadeh (2017) found that both interlocutor's status (professor vs. fellow student) and familiarity (familiar vs.

unfamiliar) influenced L2 speakers' anxiety in a short conversation, with an unfamiliar professor eliciting the highest level of anxiety.

## The Current Study

According to social psychology, people react to the world through their perceptions rather than through their direct sensory experiences such as through sight, smell, or touch (Edwards, 1999; Langer & Abelson, 1974). This is clearly illustrated in RLS research, where listeners tend to react to speech such as by rating it as less comprehensible and intelligible and by evaluating the speaker as less competent or trustworthy if they imagine the speaker to be an L2 speaker, a judgement likely based on the speaker's non-White ethnicity (Babel & Russell, 2015; Gnevsheva, 2018; Rubin, 1992; Kang & Rubin, 2009; Kutlu et al., 2022). Listeners' judgement is, in fact, governed not by the actual speech signal they receive but by their perception, which reflects the often stereotypical association between race and language that they must have internalized through prior experience (Kubota & Lin, 2006; Rampton, 1990).

Whereas numerous studies have captured various forms of RLS phenomena and have documented their effects on the perception and comprehension of a speaker's speech (Kang & Rubin, 2009; Hu & Su, 2015; Kennedy et al., 2024; Kutlu et al., 2022), less is known about how RLS could potentially affect speakers' emotional states. Therefore, this study focuses on a semi-interactive setting of an English oral exam by presenting the examiners as either Caucasian or South Asian and asking L2-speaking participants to respond to the questions asked by the examiners from these two different backgrounds. Because L2 anxiety as an emotional state is triggered by a person's perception of the speaking situation and the interlocutor (Peirce, 1995; Shirvan & Talebzadeh, 2017; Woodrow, 2006), it is important to understand how L2 speakers' perception of examiners could have an impact on their feeling of anxiety while speaking,

especially in a medium- to high-stakes situation such as an oral exam (Horwitz et al., 1986). This study is thus guided by the following research question: Is the examiner's appearance (Caucasian vs. South Asian) associated with different levels of speaking anxiety for L2 speakers completing an oral exam?

## Method

### Participants

The participants were 40 L2 English-speaking students originally from China (14 women, 26 men; $M_{age}$ = 26.4 years, *SD* = 4.64, range = 17–40), all students or recent graduates from an English-medium university in Montreal, Canada. All participants identified Mandarin Chinese as their L1, and 34 (85%) reported varying degrees of knowledge of additional languages beyond English and Mandarin, including French (32), Japanese (8), Korean and Cantonese (2 each), and Spanish (1). Before their arrival in Canada, 31 participants (77.5%) had no prior experience living or studying in an English-speaking country. Among the remaining nine, four had lived in the United States (10 months to 4 years), two in the United Kingdom (13 months, 2 years), one in Australia (3 months), one in India (5 months), and one in both the United States (2 months) and Australia (4 months). Participants reported studying at undergraduate (12), MA/MSc (13), or PhD (7) levels (with eight participants choosing not to disclose their level of study) in various disciplines, including engineering (18), computer science (5), business (6), science (5), social science (3), and arts and humanities (3). As international students, they had resided in Montreal for a mean of 3.8 years (*SD* = 3.25, range = 1 week–10.7 years) and had studied English for a mean of 13.8 years (*SD* = 4.63, range = 4–25). Of the 40 participants, 37 (92.5%) reported having taken standardized English exams, including Academic IELTS (*n* = 25, $M_{Overall}$ = 6.5, $M_{Speaking}$ = 6.0), TOEFL (*n* = 7, $M_{Overall}$ = 89, $M_{Speaking}$ = 21), CAEL (*n* = 2, $M_{Overall}$ = 65, $M_{Speaking}$ = 55), CELPIP (*n* = 1, Overall = 5.5, Speaking = 5.5), Duolingo (*n* = 1, Overall = 120, Speaking = 125), and General Training IELTS (*n* = 1, Overall = 7.5, Speaking = 6.0). Using a 9-point scale (1 = “beginner,” 9 = “nativelike”), participants self-rated their English proficiency in speaking

(*M* = 5.45, *SD* = 1.24), listening (*M* = 6.42, *SD* = 1.13), reading (*M* = 6.90, *SD* = 1.24), and writing (*M* = 6.0, *SD* = 0.88), which suggested that they had fairly high L2 skills.

## Materials

There were five sets of materials: (a) a background questionnaire; (b) the main experimental materials that included auditory and video stimuli; (c) the rating scales targeting participants' assessment of the examiners, their own performance, and their speaking anxiety at different timepoints; (d) the posttest survey eliciting participants' social categorization of the examiners as well as their familiarity with the two test topics; and (e) the instructions and prompts for the post-experimental retrospective recall interview.

### *Background Questionnaire*

The background questionnaire (see Appendix A) elicited detailed information about participants' age, gender, country of origin, education, their arrival and residence in Montreal, and their previous study-abroad experience. It also targeted participants' language background and their prior history learning and using other languages. Participants were asked to list any standardized English tests taken in the past and to report their latest overall and speaking test scores. Finally, they were also asked to rate their English proficiency level in speaking, listening, reading, and writing on a 9-point scale.

### *Experimental Materials*

**Prompt questions.** Six prompt questions were developed for the main test (see Appendix B), with three questions targeting one of the two topics: artificial intelligence (AI) and social media. The topics were selected to reflect general knowledge and everyday relevance for university-level students. Each topic was introduced through a brief opening statement, followed by three questions designed to elicit extended spoken responses. The first question in each set

asked about the advantages and disadvantages of a given situation, the second required participants to provide specific examples based on personal experience, and the third elicited an opinion with justification. This structure ensured that responses reflected a range of cognitive and discourse demands. The two sets of prompts were comparable in length (50 vs. 48 tokens) and lexical density (type–token ratios of 0.50 vs. 0.56) and had similar proportion of high-frequency vocabulary (K1: 82% vs. 79%), according to a lexical profile analysis (https://www.lextutor.ca/vp/eng).

**Audiovisual stimuli.** To create the audio stimuli for a matched-guise design, two self-identified female L1 Canadian English speakers (Speakers A and B), both born and raised in Vancouver, Canada, each recorded the six spoken prompts for the topics of AI and social media. Speaker A recorded all prompts first following detailed guidelines (e.g., professional tone, close microphone placement, quiet setting), and Speaker B then imitated her delivery as closely as possible. Although these audio-recordings were initially intended for direct pairing with the video-recordings of different individuals (see below), aligning the audio and video tracks proved difficult due to natural variation in articulatory settings and segmental realization across speakers. For example, one speaker pronounced the /w/ in *ways* as [w], while another produced it with lip movements visually compatible of a [v]-like articulation—resulting in visible mismatches between articulator movement and sound. To ensure precise audiovisual synchronization and experimental control, an AI-based voice cloning approach was therefore adopted. A clean 8-second voice sample from each speaker was extracted and used to generate two distinct voice models using CapCut (https://www.capcut.com), a multimedia editing platform with text-to-speech (TTS) synthesis capabilities. These models were then used to synthesize the six prompts in each cloned voice, resulting in 12 audio files. Most importantly,

each cloned voice belonged to the original speaker, in the sense that it retained their voice and speech quality, and was therefore clearly identifiable as a separate female speaker of Canadian English.

To create the visual component of the stimuli, two women were recruited to serve as examiners: one self-identified as South Asian (age 28) and the other as Caucasian (age 31). To minimize potential confounds such as physical attractiveness (Rubin, 1992), both were video-recorded under controlled conditions: identical room settings, plain background, consistent lighting, dark one-color clothing, neutral facial expressions, and head-and-shoulder framing (see Figure 1). Under the supervision of the researcher, each woman reviewed the six written scripts and then recorded themselves reading each prompt aloud, assuming the role of a real test examiner while looking directly into the camera. This process resulted in 12 video clips (six per examiner).

**Figure 1.** Screenshots from the video recordings of the two examiners

In post-production, the original audio was removed from each examiner's video. The corresponding AI-generated speech (from either Speaker A's or Speaker B's cloned voice) was overlaid onto each video clip using CapCut's automated lip-syncing feature. This tool adjusted

mouth movements to align with the audio track, resulting in 24 audiovisual stimuli in which each examiner produced the six prompts using either of the two cloned voices. These materials were then organized in eight sets of matched video stimuli, each with a different visual examiner guise (Caucasian vs. South Asian), topic (AI vs. social media), and actual audio track (Speaker A vs. Speaker B), as summarized in Table 1. The total length of videos in each condition ($M_{length}$ = 17.5 s, *SD* = 1.20, range = 16–19) was within a typical duration range of the stimuli used for evaluation of L2 speech (Nagle et al., 2019; Saito et al., 2016; Uchihara et al., 2023).

**Table 1**. *Video Stimuli*

| Video set | Examiner | Topic | Audio track |
|---|---|---|---|
| 1 | Caucasian | AI (3 questions) | Speaker A (L1 English) |
| 2 | Caucasian | AI (3 questions) | Speaker B (L1 English) |
| 3 | Caucasian | Social media (3 questions) | Speaker A (L1 English) |
| 4 | Caucasian | Social media (3 questions) | Speaker B (L1 English) |
| 5 | South Asian | AI (3 questions) | Speaker A (L1 English) |
| 6 | South Asian | AI (3 questions) | Speaker B (L1 English) |
| 7 | South Asian | Social media (3 questions) | Speaker A (L1 English) |
| 8 | South Asian | Social media (3 questions) | Speaker B (L1 English) |

To establish that the videos were matched in approximate level of content clarity and difficulty as well as video quality, 12 L2 English-speaking Chinese international students (10 women, 2 men; $M_{age}$ = 30.4 years, *SD* = 7.74, range = 19–45) were recruited to pre-rate the videos. All were students from China pursuing undergraduate (3), MA (5), or PhD (4) degrees in different Canadian institutions, and they reported either Mandarin Chinese (11) or Cantonese (1)

as their L1. As international students, the pre-raters all had prior experience taking a standardized English test, including IELTS (9), TOEFL (2) and CELPIP (1); therefore, all were familiar with the general assessment context of this study.

The pre-raters evaluated all videos, where three pre-raters were assigned to a subset of six videos featuring one examiner–voice pairing (e.g., Caucasian examiner paired with Speaker A's voice), with three prompt questions targeting the topic of AI and three targeting the topic of social media. Using Qualtrics (https://www.qualtrics.com), they used 7-point scales to read each prompt question and evaluate its clarity (1 = "very unclear," 7 = "very clear") and difficulty (1 = "very difficult," 7 = "very easy"). They then watched each video to assess its quality (1 = "very poor quality," 7 = "very good quality"). Because the pre-rating data violated the assumption of normality, as demonstrated visually through inspection of Q–Q plots and statistically through significant Shapiro-Wilk's tests, $W > .75$, $p < .006$, these data were analyzed non-parametrically. According to Wilcoxon tests (one-tailed), the two test sets focusing on AI and social media (SM) were perceived as comparable in prompt clarity, $W = 0.01$, $p = .993$ ($Mdn_{AI} = 6.33$ vs. $Mdn_{SM} = 6.50$) and prompt difficulty, $W = 25.00$, $p = .776$ ($Mdn_{AI} = 5.67$ vs. $Mdn_{SM} = 6.00$), as well as in video quality, $W = 4.50$, $p = .828$ ($Mdn_{AI} = 6.50$ vs. $Mdn_{SM} = 6.42$).

All finalized target video clips were then arranged in eight test conditions in Qualtrics by creating all possible combinations of the examiner's identity (South Asian vs. Caucasian), prompt order (AI vs. social media), and the speaker's audio embedded in the video clips (Speaker A vs. Speaker B). To counterbalance the order effect, four conditions featured the Caucasian examiner in the first test and four featured the South Asian examiner in the first test, meaning that across all conditions there was always a switch in examiner identity from South

Asian to Caucasian or from Caucasian to Asian between the first and the second test set. These conditions are summarized in Table 2.

**Table 2**. *Test Conditions*

| | Test 1 | | | Test 2 | | |
|---|---|---|---|---|---|---|
| Condition | Examiner | Topic | Audio | Examiner | Topic | Audio |
| 1 | Caucasian | AI | Speaker A | South Asian | Social media | Speaker B |
| 2 | Caucasian | Social media | Speaker A | South Asian | AI | Speaker B |
| 3 | South Asian | AI | Speaker A | Caucasian | Social media | Speaker B |
| 4 | South Asian | Social media | Speaker A | Caucasian | AI | Speaker B |
| 5 | Caucasian | AI | Speaker B | South Asian | Social media | Speaker A |
| 6 | Caucasian | Social media | Speaker B | South Asian | AI | Speaker A |
| 7 | South Asian | AI | Speaker B | Caucasian | Social media | Speaker A |
| 8 | South Asian | Social media | Speaker B | Caucasian | AI | Speaker A |

A pilot run of the experimental materials was conducted with two students recruited from the same participant pool (i.e., Chinese international students with English as their L2 studying in Montreal, Canada). Based on the feedback gathered from the two pilot participants, it was decided to set the time limit for participants to answer each prompt question at 45 seconds, with 5 seconds of preparation time. This timeframe allowed participants to express ideas while feeling sufficient time pressure imposed by the time limit. The pilot participants also verified that the instructions were clear and ensured that the testing platform was easy to navigate.

### *Rating Scales*

The target measures included three sets of 100-point sliding scales (provided in Appendix C). The first set targeted participants' self-perceived degree of speaking anxiety (1 = "not anxious at all," 100 = "very anxious"). Participants each rated their anxiety before and after each of the two tests that they completed under different examiner guises, resulting in four anxiety ratings per participant. The second set of measurement (adapted from McDonough et al., 2022) included several scales eliciting participants' perceptions about each examiner's comprehensibility (i.e., whether the examiner is easy to understand, 1 = "very easy to understand," 100 = "very difficult to understand"), accentedness (i.e., whether the examiner is heavily accented, 1 = "not accented at all," 100 = "heavily accented"), and participants' willingness to have the examiner in a real IELTS speaking test (1 = "not at all," 100 = "very much"). An additional scale set targeted participants' self-assessment of their own speaking performance (1 = "very poorly," 100 = "very well") before and after each test, capturing anticipated and actual perception of their speaking performance; however, these ratings are not analyzed further because they fall outside the immediate scope of this study. At the end of each test, there was a comment section, where participants could leave a brief comment and share their experience and feelings regarding each test.

### *Posttest Survey*

Because RLS importantly depends on people's social categorization of a speaker into a social group such as by their ethnicity, a posttest survey was used to gather this information (see Appendix D). The survey included a still image of each examiner, followed by a multiple-choice question asking participants to identify the examiner's racial or ethnic background from seven categories: East Asian, Southeast Asian, South Asian, Latin/Central/South American, Middle

Eastern, European, or African. Participants then rated the perceived nativeness of each examiner's English on a 100-point sliding scale (1 = "second language speaker of English," 100 = "native speaker of English"). Lastly, to account for potential topic effects, participants were asked to estimate their familiarity with each of the two topics—AI and social media—on separate scales (1 = "not at all familiar," 100 = "very familiar").

***Retrospective Recall***

To gain an insight into participants' experience taking the test under different examiner guises and to determine how this experience relates to their speaking anxiety, a semi-structured retrospective recall interview (see Appendix F) was carried out with eight participants, whose anxiety ratings before the two tests taken under different visual guises showed a relatively big difference (i.e., with an anxiety gap between the Asian and the Caucasian examiner guises equalling or exceeding 20 points on a 100-point scale). All eight interviews were conducted in Mandarin Chinese, the shared L1 between the researcher and participants. To understand the overall emotional status of each participant, two initial questions were asked (i.e., "Overall, how do you feel about the two speaking tests? *对于两个口语测试，你的整体感觉怎么样*?" and "Was there a specific moment during the whole test that made you particularly nervous? *你有没有在测试中的什么时候感觉到特别紧张*?"). Subsequently, the screen recording of the participant's completion of both tests was played back to them to help focus their recall on specific moments during each test that were of particular significance to them. Participants were encouraged to pause the video at any point where they felt something was worth discussing, allowing them to take the lead in identifying emotionally or cognitively salient moments. During the playback, each participant was also asked questions probing their justification for each of their ratings (e.g., "Your anxiety before test was X, could you explain why it was so high/low?

*你在测试前的焦虑值是X，你可以解释一下为什么它这么高/低吗?*"), with additional follow-up questions or clarification requests when needed (e.g., "Do you think your anxiety affected your performance? *你认为你的焦虑值影响你的表现了吗?*" or "What was it about the examiner that made you anxious? *考官的哪一方面让你感到焦虑?*"). Because only the participants with a considerable anxiety gap between the two visual guises were interviewed, there was a final question asking specifically about this gap (i.e., "Your anxiety before the first test was X, and it was Y before the second test. Can you explain why there is a relatively big gap between the two ratings? *你的焦虑值在第一个测试前是X，在第二个测试前是Y，你能解释一下为什么这两个数值差距比较大吗?*").

**Procedure**

Participants first contacted the researcher via email or social media to reserve a one-hour individual session in a quiet computer-equipped room. Participants were informed beforehand that they would be participating in two short English speaking tests and that they would be evaluating their experience taking those tests. Forty participants were then randomly assigned to one of the eight matched-guise conditions before their arrival (as shown in Table 2), with five participants per condition. When participants arrived for the testing session, they first read and signed the consent form. The researcher then sat next to participants to introduce the nature and content of the speaking tests (i.e., the sequence of each task and the approximate length of each). At this point, a digital audio-recorder was turned on with consent from participants, and all participants were required to set their phone to silent mode for the entire testing session.

After receiving general instructions, participants then used a practice test to familiarize themselves with the testing procedure, while the researcher sat next to them to provide guidance and clear up any confusion. Because the practice test was designed solely to acquaint participants

with the task structure and the testing platform prior to the actual test, only one prompt question (also listed in Appendix B) was used, covering an unrelated topic (i.e., making friends). The question mirrored the format of the first question from the actual test sets (i.e., asking about advantages and disadvantages), but instead of using a clear visual of an examiner, it featured a darkened silhouette of a woman's upper body and head (see Figure 2). Similarly, the audio track adopted for the practice test was different from those used in the actual experimental materials; instead, it was recorded by an L1 speaker of American English. During the practice test, participants self-navigated through the testing platform, and practiced responding to the practice question.

**Figure 2.** Silhouette image used in the practice test

After clarifying all remaining questions that emerged during the practice test, the researcher then initiated the relevant test condition in the Qualtrics interface, reminded participants to take a break after the first test, started screen recording, and then left the room to allow participants to work individually. At the beginning of each test, participants first saw a static image of the first examiner and rated their anxiety level, as triggered by the appearance of the examiner. Participants then played the three video clips in the first test, listening to each

prompt question and responding to each by following on-screen instructions (see Appendix E). In all cases, each question (i.e., a separate video clip) could only be played once, with participants given 5 s to prepare their answer and 45 s to record their answer before hearing the cut-off beep sound (which was the optimal response timeframe determined through pilot testing).

Once participants finished their first test (i.e., responding separately to the three prompts by the first examiner), they then rated their speaking anxiety for the second time, estimating their anxiety immediately after completing the test. They also evaluated the examiner's comprehensibility and accentedness and their willingness to have the examiner in a real IELTS test. At this point, participants were also encouraged to provide comments about their experience with the test. Before taking the second test (with a change of examiner guise, voice, and topic), participants took a short break (approximately 5–8 minutes) during which they filled out the background questionnaire. To minimize external influence, the researcher did not interact with participants at this point beyond assisting with the background questionnaire. After collecting the background questionnaire, the researcher left the room again so that participants could continue the second test individually, following the same procedure. Altogether, each participant provided audio responses to six prompts about two topics as part of two tests, each featuring the voice of an L1 Canadian English speaker but under a different visual guise (Caucasian vs. South Asian examiner).

The researcher entered the room again when participants finished the second test. At this point, participants answered the posttest survey on a different laptop (see Appendix D) while the researcher checked participants' anxiety scores to determine if an audio-recorded retrospective recall session should follow. Those who were selected to participate in the recall interview (which generally followed the structure described in Appendix F) first signed a separate consent

form and then engaged in a semi-structured discussion with the researcher while watching the screen recording of their tests. At the end of session, all participants were debriefed about the full purpose of the study, which focused on the use of different videos to investigate a test-taker's speaking anxiety triggered by examiners of different ethnicities. All participants agreed not disclose the study goal to other prospective participants, and signed another consent form allowing the use of their data before leaving the testing session.

## Data Analysis

The study employed both quantitative and qualitative analyses to examine whether the perceived racial identity of English test examiners contributed to participants' speaking anxiety and their evaluations of examiner speech. The quantitative data were exported from Qualtrics to spreadsheets and analyzed in Jamovi (The Jamovi Project, 2025). The ratings were first checked for internal consistency (two-way random average agreement intraclass correlations), which showed acceptable values for anxiety (.91), with four ratings per participant (two in each test), and for comprehensibility (.83) and accentedness (.73), each with two ratings per participant (one in each test). In contrast, internal consistency was low for the rated willingness to have the examiner in the future (.36), with two ratings per participant (one in each test), implying that participants disagreed across the two tests in their judgment about the examiner. Considering that all ratings were meant to be treated separately (i.e., without computing their averages per participant), all consistency values were deemed sufficient.

The quantitative data were also checked for various statistical assumptions. The anxiety ratings were normally distributed, with Levene's tests of homogeneity of variances revealing nonsignificant values in Test 1, $F(1, 38) < 0.16$, $p > .693$, and Test 2, $F(1, 38) < 0.01$, $p > .951$, and the sphericity assumption met in a repeated-measures design. Therefore, the anxiety ratings,

which passed all checks, were analyzed through mixed ANOVAs. In contrast, according to normality checks (Shapiro-Wilk), $W > 0.73$, $p < .028$, and visual inspection (Q–Q plots), the ratings of accentedness and comprehensibility demonstrated a positive skew whereas the rated willingness to have the examiner in the future as well as the nativeness of the examiner's English and participants' familiarity with the test topics (both evaluated after the speaking tests) showed a negative skew. Therefore, all analyses involving these ratings were carried out through non-parametric procedures. For all statistical analyses, the alpha level for significance was set at .05 and was Bonferroni-adjusted for multiple comparisons. Effect sizes were interpreted using partial $\eta^2$ (.01, .06, and .14) for ANOVA effects (Richardson, 2011), Cohen's *d* (0.40, 0.70, and 1.00) for pairwise comparisons using parametric tests, and *r* (.25, .40, and .60) for pairwise comparisons using nonparametric tests (Plonsky & Oswald, 2014), where each value designates small, medium, and large effects, respectively.

To complement the statistical findings, qualitative data were drawn from retrospective recall interviews conducted after the speaking tasks. Eight interviews were transcribed and automatically translated from Mandarin Chinese to English using Turboscribe (https://turboscribe.ai). To ensure both accuracy and completeness, each transcript was manually reviewed by the researcher. This involved listening to the original Chinese audio recordings in full, checking the accuracy of the English translation and revising any misinterpretations, omissions, or unclear wordings. Particular attention was paid to preserving the intended meaning of participants' responses, especially in the segments cited to illustrate key points in the discussion.

Because the qualitative component of this study targeted a selective group of participants (i.e., those with a large discrepancy in anxiety ratings between the two test conditions) and

focused on their subjective experiences during the speaking tasks—particularly their emotional responses and perceptions of the two examiner guises—a qualitative descriptive approach was adopted (Sandelowski, 2000; Thorne et al., 1997). Rather than applying a formal coding scheme or conducting a full thematic analysis, participants' comments were used to contextualize and interpret key quantitative patterns (e.g., high pretest anxiety when facing the South Asian examiner in Test 2), offering what Sandelowski (2000, p. 336) describes as a "comprehensive summary of events in the everyday terms of those events." The researcher first reviewed all eight transcripts to identify participant comments relating to emotional responses, perceptions of examiner speech or appearance, and self-rated anxiety scores. A second, targeted pass through the transcripts was conducted to extract comments that directly helped explain specific quantitative trends, such as the pretest anxiety gap between the examiner guises and the observed decrease in anxiety with the South Asian examiner in Test 2. This approach enabled the integration of statistical and descriptive evidence while preserving the participants' perspectives reported in their own words.

## Results

### Preliminary Analyses

In order to answer the research question, which asked whether the examiner's visual identity (Caucasian vs. South Asian) is associated with different levels of speaking anxiety for L2 speakers completing an oral exam, it was essential to first establish participants' social categorization of the two examiners. Therefore, the first analysis focused on the results from the posttest survey asking participants to select the examiner's racial background from seven categories: East Asian, Southeast Asian, South Asian, Latin/Central/South American, Middle Eastern, European, or African. Participants categorized the Caucasian speaker as European (60%), Latin/Central/South American (25%), or Middle Eastern (15%). They perceived the Asian speaker as belonging to several categories, including South Asian (63%), Latin/Central/South American (18%), African (10%), Southeast Asian (7%), and Middle Eastern (2%). Nevertheless, in both cases, approximately the same number of participants correctly identified the examiner's ethnic origin (about 60% or 24 of the 40 participants), and there were no responses placing each examiner in the other's racial category. Even though both examiners' audio was recorded by L1 speakers, participants also provided significantly different ratings for the nativeness of the examiners' English in the posttest survey. According to a Wilcoxon test, they considered the Caucasian examiner's English to be more nativelike than the South Asian examiner's English ($Mdn_{\text{Caucasian}} = 84.50$ vs. $Mdn_{\text{Asian}} = 68.50$, where 100 meant "native speaker of English"), $W = 514.50$, $p < .001$. Both these findings suggested that the visual manipulation was broadly effective in conveying the examiner's intended racial identity and at activating stereotypes that Asian-looking examiners speak a less standard (less native) variety of English.

Second, before analyzing participants' speaking anxiety, it was also important to ensure that they were similarly familiar with the two test topics, as topic familiarity—and therefore any potential response uncertainty associated with less familiar or expected content—may have influenced their affective status and performance during the test. In the posttest survey, participants reported comparable levels of familiarity with the topics of AI (*Mdn* = 79, range = 1–100) and social media (*Mdn* = 67, range = 20–100), with no difference in the ratings according to a Wilcoxon test, $W = 337.50$, $p = .499$, suggesting that both topics were similarly accessible to them. This implied that any potential differences in speaking anxiety are unlikely to have arisen due to unequal topic familiarity.

Finally, it was essential to examine whether there were any differences in participants' response patterns as they answered each question prompt. With three exceptions, participants used the entire allotted time (45 seconds) to respond to each prompt. The exceptions included one participant going over the allotted time limit for all three questions in Test 1, a different participant skipping one question in Test 1 because they did not understand the prompt, and another participant accidentally skipping one question in Test 2. Although no comprehensive analysis of response quality has been carried out at this stage, a preliminary check of response length (i.e., total words spoken in response to each question prompt) indicated that participants produced between a mean of 76 and 86 words in response to the three questions in Test 1 and between 68 and 75 words in response to the three questions in Test 2 (see Appendix G for full descriptive statistics). Most importantly, response length did not differ as a function of the examiner's guise (Asian vs. Caucasian) for any of the questions, as shown by Wilcoxon tests for responses in Test 1, $W > 153.00$, $p > .208$, and in Test 2, $W > 138.00$, $p > .147$.

**Speaking Anxiety**

Before Test 1 and Test 2, participants' anxiety levels (see Table 3) fell on average within the midrange of the scale, approximately between the ratings of 40 and 60 on a 100-point scale. More importantly, there was variability in anxiety levels reported, with participants using the entire scale range to estimate their anxiety.

**Table 3.** *Descriptive Statistics for Anxiety Ratings*

| | Before test | | | After test | | |
|---|---|---|---|---|---|---|
| Examiner | *M* | *SD* | Range | *M* | *SD* | Range |
| | | | Test 1 | | | |
| Asian | 44.70 | 24.33 | 1–81 | 44.35 | 24.42 | 0–94 |
| Caucasian | 54.25 | 26.92 | 0–90 | 57.45 | 23.25 | 10–100 |
| | | | Test 2 | | | |
| Asian | 60.45 | 24.45 | 19–100 | 46.50 | 24.99 | 9–100 |
| Caucasian | 40.90 | 24.40 | 0–82 | 41.80 | 24.62 | 0–90 |

To examine whether participants' anxiety in Test 1 differed as a function of the interviewer's visual guise, we carried out a mixed two-way ANOVA with visual guise (Asian vs. Caucasian) as a between-participants variable and time (before vs. after test) as a within-participants variable. The ANOVA revealed no statistically significant effects for visual guise, $F(1, 38) = 2.37$, $p = .132$, partial $\eta^2 = .06$ (weak effect), time, $F(1, 38) = 0.28$, $p = .602$, partial $\eta^2 = .01$ (weak effect), or a two-way interaction, $F(1, 38) = 0.43$, $p = .516$, partial $\eta^2 = .01$ (weak effect). These results (illustrated in Figure 3) suggested that participants did not differ in their perceived anxiety before or after Test 1 as a function of the interviewer's visual guise.

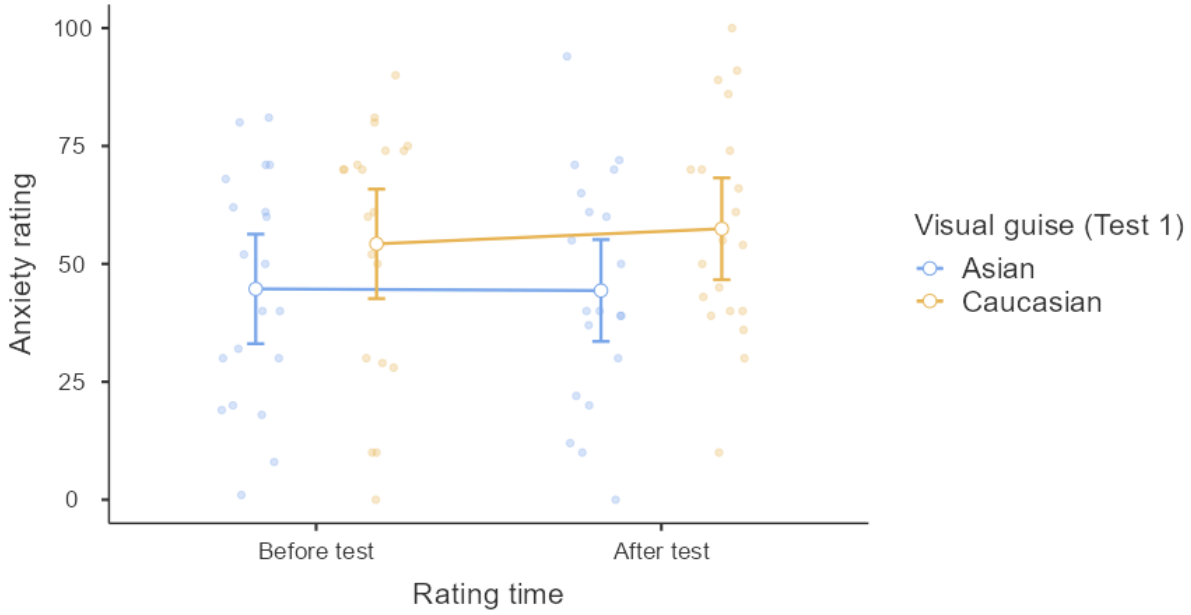



**Figure 3**. Participants' anxiety in Test 1 plotted as a function of interviewer's visual guise (Asian vs. Caucasian) and rating time (before vs. after test). Error bars enclose a 95% confidence interval (CI).

To examine participants' anxiety in Test 2, we carried out a similar mixed two-way ANOVA. The ANOVA revealed a statistically significant effect for time, $F(1, 38) = 5.44$, $p = .025$, partial $\eta^2 = .13$ (medium effect), and a statistically significant two-way interaction, $F(1, 38) = 7.04$, $p = .012$, partial $\eta^2 = .16$ (large effect), but no significant main effect of visual guise, $F(1, 38) = 2.79$, $p = .103$, partial $\eta^2 = .07$ (medium effect). According to Bonferroni-corrected post hoc tests exploring the significant interaction (illustrated in Figure 4), before starting the test, participants estimated their anxiety higher after seeing the Asian than Caucasian interviewer, $t(38) = 2.53$, $p = .016$, $d = 0.80$ (medium effect), although this difference missed statistical significance based on the adjusted alpha level of .0125. After completing the test,

participants did not differ in their anxiety between the two visual guises, $t(38) = 0.60$, $p = .553$, $d = 0.19$ (weak effect), which reflected a drop in perceived anxiety reported by participants in the South Asian guise condition pre-to-posttest, $t(38) = 3.53$, $p = .001$, $d = 0.59$ (medium effect).

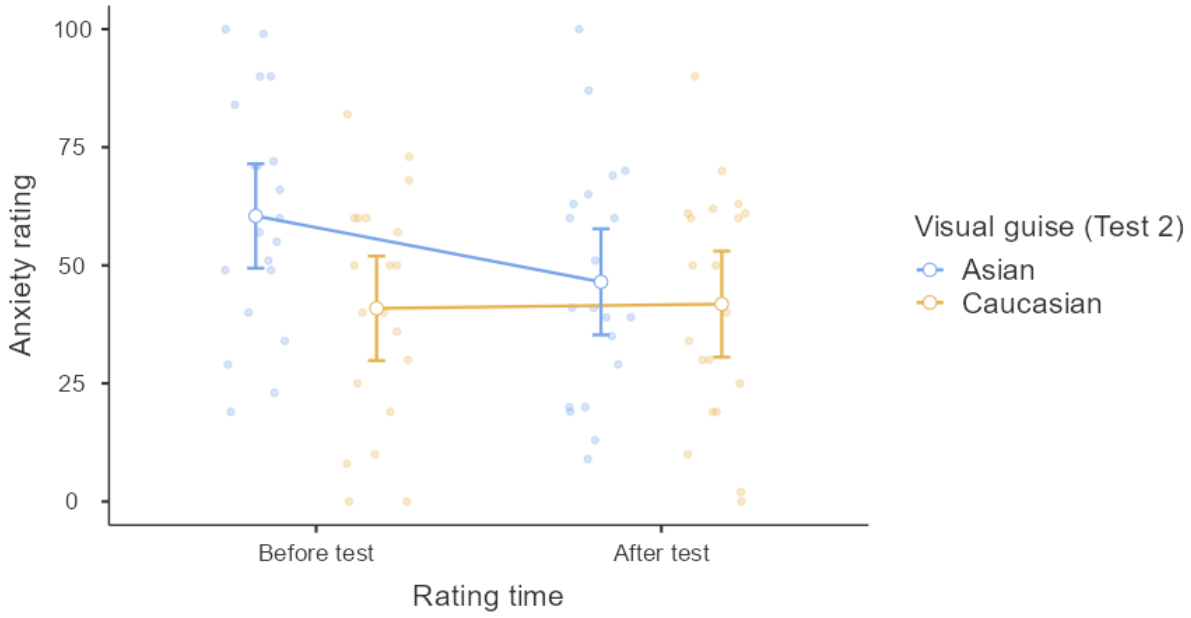



**Figure 4**. Participants' anxiety in Test 2 plotted as a function of interviewer's visual guise (Asian vs. Caucasian) and rating time (before vs. after test). Error bars enclose a 95% confidence interval (CI).

The analysis of retrospective recall interviews provided further details about the potential relationship between participants' reported anxiety and the interviewer's appearance, with four of the eight interviewees explicitly mentioning this link. For example, when seeing the image of the South Asian examiner, P14 stated: "From the moment I saw her, my mind went blank." Similarly, when asked to identify any particular moment when they felt nervous, P21 pointed out: "I was a bit nervous when I saw that picture of an Indian [examiner]." When asked to

explain the anxiety rating gap between the Caucasian (28) and the South Asian (71) examiners, P22 answered briefly: “Appearance.” P5 provided an extended explanation for why the Asian examiner’s appearance was associated with heightened anxiety for them, associating the examiner’s ethnicity with their assumption about the questions that the examiner might ask:

> I was mainly afraid of the second examiner. The first examiner she looks like a European, and I thought the questions she asks may be more familiar. Because I usually watch some Western stuff. I am more familiar with Western culture. The second [examiner] I think she is South Asian, so she might ask some questions about their culture background. I know this may not actually be true, but I have this subconsciousness that makes me think she might ask me something about her cultural background, and I thought that would be difficult to answer. I know the question will be random, but I have this subconsciousness.

For P5, their assumption made them feel that they need to “be careful” and “pull [themselves] together to deal with” the examiner, which reflected their anxiety rating before the test. However, the Caucasian examiner’s appearance also elicited anxiety for this participant, although for different reasons:

> The reason I chose 60 [to rate my anxiety when seeing the Caucasian examiner] is because I think she might be a native speaker of English… I feel a bit nervous… I think she looks like a Westerner, European or Latino that speaks English. Because some of my errors may not be a problem, but for those native speakers it might be a glaring mistake. The moment they hear it, they know something is off, or very glaring… It’s like if a foreigner says a very weird word in Chinese, and then you feel it’s very jarring. This kind of feeling.

When asked to elaborate, P14 attributed their anxiety to the stereotypical association between a speaker's appearance and comprehensibility: "I think when I saw the second examiner [South Asian], I naturally felt the anxiety, because I got worried. I have the preconception, the stereotype, that I might [get] stuck because I could not understand." P22, whose pretest anxiety was 71 for the South Asian examiner and 28 for the Caucasian examiner, attributed their feelings to their prior negative experience with understanding an interlocutor: "Because I have a group activity this year, there was a girl who looks very much like her [the South Asian examiner], 80% like her, and then, I felt very helpless." Although this particular comment illustrated a specific real-life experience, other participants appeared to draw on second-hand information, likely informed by preconceived ideas about individuals from South Asian ethnic backgrounds, as explained by P21: "Because I read a lot of news and posts by others online about Indian examiners and their accent being a problem."

Some other factors besides examiner appearance were also reported, with relevance to participants' speaking anxiety. First, for some participants, speaking in the L2 itself is anxiety-provoking, as illustrated by P9: "I feel that to us, our English is a second language, [so] it takes some effort from our brain when you speak English. So it's definitely a little bit anxiety-provoking." Similarly, P31 commented: "If this is a Chinese test, I might not be anxious at all. But this is not." Second, the testing environment could also evoke the feelings of tension and stress, as explained by P14: "It's like this, for this kind of exam, it requires logic and consistency. This is what I'm lacking. [The test] brought me back directly to the state of taking the IELTS exam. I have analyzed this situation before. I'm going to have an entire breakdown. Every time after the test, I need a long time to recover." Such tension, experienced before the test, could result in continued feelings of anxiety, as pointed out by two participants:

- Basically, I am like, if I relaxed in the beginning, then the rest might be alright. It depends on the degree of difficulty. But I got nervous at the beginning, it's already guaranteed it won't be good later. (P14)
- During the practice test, it really scared me because it was a short and simple question, but I didn't understand it, so it made me anxious… Because I thought that [the next test] might be the same situation. (P31)

Participants' lack of knowledge to answer questions also appeared to be anxiety-provoking for some, causing them to "freeze," "make up answers" (P29), and experience "a lot of blank moments" (P31). Although most participants described anxiety in a negative way, P29 also pointed out its facilitative role, particularly in helping with test performance: "I think I had to make myself nervous. Otherwise, when I answer the question, I would not perform well. So I thought I should be more nervous."

**Evaluations of Examiner**

At the end of each test, participants provided ratings for each examiner targeting three dimensions: accentedness, comprehensibility, and willingness to have the examiner in a real speaking test (see Table 4). Generally speaking, participants evaluated both examiners as not very accented (with average ratings of about 10–30 on a 100-point scale, where 1 meant "not accented at all"), easy to understand (with average ratings of about 15–20, where 1 meant "very easy to understand"), and were willing to have them as interviewers in a real exam situation (with average ratings of about 70–80, where 100 meant "very much"). However, there was a large amount of variability in participants' responses, where they used the entire scale range to evaluate the examiners.

**Table 4.** *Descriptive Statistics for Examiner Evaluations*

| | Test 1 | | | Test 2 | | |
|---|---|---|---|---|---|---|
| Interviewer | *M* | *SD* | Range | *M* | *SD* | Range |
| | | | Accentedness | | | |
| Asian | 28.6 | 29.9 | 0–97 | 11.8 | 10.7 | 0–33 |
| Caucasian | 19.6 | 23.5 | 0–84 | 16.9 | 23.8 | 0–93 |
| | | | Comprehensibility | | | |
| Asian | 15.3 | 24.6 | 0–99 | 14.9 | 18.8 | 0–74 |
| Caucasian | 21.0 | 26.1 | 0–97 | 16.1 | 24.6 | 0–100 |
| | | | Willingness to have interviewer as examiner | | | |
| Asian | 69.5 | 30.5 | 0–100 | 74.6 | 19.3 | 22–100 |
| Caucasian | 74.8 | 21.1 | 6–100 | 78.7 | 17.1 | 50–100 |

These ratings were compared through nonparametric Mann-Whitney U tests to determine whether the examiners elicited different evaluations as a function of their visual identity. In Test 1, no significant differences emerged between the evaluations of the South Asian and the Caucasian examiner in accentedness, $U = 168.50$, $p = .399$, $r = .16$, comprehensibility, $U = 164.50$, $p = .337$, $r = .18$, or willingness to have the examiner in a real test, $U = 185.50$, $p = .704$, $r = .07$. Similarly, in Test 2, no significant differences were found for examiner accentedness, $U = 194.00$, $p = .881$, $r = .03$, comprehensibility, $U = 198.50$, $p = .978$, $r = .01$, or willingness to encounter the examiner in a real test, $U = 181.50$, $p = .626$, $r = .09$. Thus, whereas participants' evaluations varied slightly across the two examiners, participants did not attribute different ratings to examiners as a function of their visual identity.

This finding was also generally supported through participants' comments in the recall interview and their brief notes provided in the comment boxes. For example, after the first test with the South Asian examiner, P12 noted: "The test is good, and the examiner's accent is clear and easy to understand." In the recall session, P22 mentioned: "Both of them don't have an accent," and P29 pointed out the consistency in both examiner's speech: "There was not much difference. In terms of accent, I think they can both be examiners. Because they speak very clearly." Even though they may have expected to hear more accented or less comprehensible speech from the South Asian examiner, several participants were in fact surprised to hear the speaker's actual speech:

- But I actually think she doesn't seem to have an accent. Even though she looks like she has an accent. But she doesn't sound like she has an accent. (P9)
- Her pronunciation is very clear, almost nothing. I thought she would have an accent, but she actually doesn't. (P14)
- I thought that Indian would speak Indian English, but she spoke quite clearly. (P22)

Nonetheless, some comments still suggested that participants' perceptions of examiners may have been shaped by subtle impressions related to the examiner's ethnicity. For example, during the replay of test videos, P21 claimed that they heard an Indian accent: "I didn't notice it when I first heard it. But now I can tell that it's a little bit of an Indian accent." P31, who rated the South Asian examiner's accent at 70 as opposed to 0 for the Caucasian examiner, believed that she detected "too heavy" an accent when taking the test with the South Asian examiner: "There's a certain accent, and it's too heavy. Even though she speaks fluently, I don't feel like it's her first language." Similarly, some participants related the Caucasian examiner's lack of accent to her appearance as well, as illustrated by P9 during the recall interview: "I don't know

how to evaluate it [accentedness]. But the overall impression is that she looks like she doesn't have an accent."

Regarding their willingness to have the examiner in a future speaking test, for some participants, their rating was influenced by the examiner's presumed linguistic performance. P31, who heard "too heavy" an accent from the South Asian examiner, provided the extreme ratings for their willingness to have the South Asian (0) versus the Caucasian (100) examiner in the future. When asked why they rejected the South Asian examiner, the participant responded: "I feel like an exam such as IELTS or TOEFL is an exam of a mainstream English-speaking country, and if it's all like that, it's weird. I feel like English exam such as IELTS and TOEFL should have a standard accent, or something traditional for the exam." Their preference for a "standard accent" was also reflected in their support of the Caucasian examiner: "First I can't hear her accent, and I think she's a native speaker. And second, she's very clear in every word, so if I meet her in the exam, I'll be very happy." A similar justification was provided by P9, whose ratings were 82 for the Caucasian examiner and 40 for the South Asian examiner: "First she [the South Asian examiner] looks mean, and then she's not like the first examiner, who... how do I put this? She doesn't look like the IELTS examiners I've had." For P22, whose rated willingness to have the Caucasian examiner in a real test was 46 points higher than their rating for the South Asian examiner (68 vs. 22), the concern about not being able to understand the examiner overpowered the actual evidence in the examiner's speech. Although P22 described the South Asian examiner as speaking "quite clearly" and sounding "quite standard" in terms of her accent, the association between the examiner's appearance and her presumed language quality contributed to their rating: "I like both of them… the second one, I rated it [willingness] very low

at first, because I was afraid that she would speak that kind of Indian English." In response to the researcher's follow-up questions, the same participant later elaborated:

P22: Because judging by her face, I think mostly people who look like this speak unclearly.

Researcher: So in a real exam you want...

P22: White people.

Researcher: Ok, because you think they will speak more clearly?

P22: Yes, more nativelike.

Researcher: But you just said, when you were listening to her [the South Asian examiner], you didn't think there's any problem with her English. But here her appearance still affects you, that you don't want her in a real exam?

P22: Yes.

Other participants, however, expressed willingness to have either examiner in a real test. For example, P29 evaluated both examiners as being easy to understand and having little accent: "They can both be examiners. Because they speak very clearly" and "I think she [the South Asian examiner] can definitely be a real examiner in terms of accent." Similarly, P5 provided a comparable willingness rating for both examiners (90 vs. 80), citing their clear pronunciation and nonaccented speech, despite initially feeling somewhat nervous about the South Asian examiner: "She [the South Asian examiner] didn't have an accent, and the words she used were easy to understand." Although P14 expressed high anxiety after seeing the image of the South Asian examiner before test—likely due to their previous negative test experience—neither their actual perception of the examiner's speech nor their willingness to have her in the future was affected:

P14: I have prior experience with an Indian [examiner]; it took me a lot of effort to just understand them.

Researcher: You have had an Indian examiner in a real IELTS exam?

P14: Yes. There have been Indian examiners and also Japanese examiners. This examiner already has a very standard pronunciation.

Researcher: (*Screen showing their rating of 90 for the South Asian examiner*) So you think she can be an IELTS examiner?

P14: Yes.

## Discussion

The goal of this research was to investigate whether reverse linguistic stereotyping (RLS), which stems from people's preconceived ideas about a person's presumed group membership, affects L2 English speakers' speaking anxiety during an oral exam. Tested through a matched-guise design (Lambert et al., 1960), participants completed two speaking tasks with video prompts featuring examiners of different racial appearances (Caucasian vs. South Asian), while the audio remained constant across conditions (i.e., Canadian English). Quantitative results showed no significant difference in speaking anxiety in the first test. However, in the second test, participants reported higher pre-test anxiety when the examiner was a South Asian woman, though this effect diminished after participants finished the test. Participants' evaluations of the two examiners—in terms of their accentedness, comprehensibility, and acceptability as real test examiners—did not significantly differ across the two visual guises, though the responses varied substantially across individuals.

Qualitative data from retrospective recall interviews provided further insight into participants' emotional responses during the tests and their evaluation of the examiners. Some participants reported initial anxiety upon seeing the image of the South Asian examiner, linking her appearance to assumptions about accented or unclear speech, cultural unfamiliarity, and prior negative experiences. Others expressed anxious feelings after seeing the Caucasian examiner, likely due to assumptions of higher language expectations and lower tolerance for errors. These findings suggest that racialized visual cues are able to activate deeply embedded social stereotypes, which in turn contribute to L2 speakers' emotional state before speaking, demonstrating that RLS could affect not only listeners' perception of speech (e.g., Hu & Su, 2015; Kennedy et al., 2024; Kutlu et al., 2022; Rubin, 1992), but also speakers' affective

emotional experiences in high-stakes testing situations. In general, the effects of RLS on speaking anxiety, as revealed through participant's self-rated anxiety and their comments from retrospective recall interviews, were marked by subtle distinctions rather than a uniform trend.

**RLS and Speakers' Self-Rated Anxiety**

First, when participants' anxiety was assessed in the first of the two consecutive tests, no statistically significant difference in speaking anxiety between the two examiner guises (Caucasian vs. South Asian) was observed, a finding consistent across both pretest and posttest anxiety ratings. One possible explanation is that participants had not yet encountered a second examiner, and thus lacked a point of comparison to activate racialized expectations. With no contrasting reference, participants may have been more focused on understanding the task itself or managing general test anxiety. In other words, their initial emotional responses may have been shaped more by task novelty than by the examiner's appearance.

This interpretation is supported by comments from several participants who described uncertainty about the test format as their primary source of anxiety. For instance, in the recall session, P29 commented that their initial nervousness stemmed from not knowing what to expect: "Of course I was a little nervous. Because I didn't expect it at first. I didn't expect [the test] to ask a more open question then let you answer it yourself." P29 later elaborated on how the anxiety gradually escalated, leading to a high rating (70) before Test 1: "Before the practice test, I wasn't so nervous. When I just came in, I was at the most relaxed, then after signing the forms and listening to your introduction, I became more and more nervous. Then after the practice test I felt that the questions were so vague and started to worry about what to say. Then I became very nervous." Similarly, P31 attributed their pretest anxiety to unfamiliarity with the test format: "Because you gave me that practice at the beginning, I didn't quite understand it. Oh,

yes, and then I think it may also be the first question. I didn't quite know [in] what form it will be given, so I was a little bit anxious." Taken together, these comments suggest that for some L2 speakers, the "test anxiety" described by Horwitz et al. (1986)—that is, the anxiety associated with oral communication under evaluation in a novel or high-stakes context—may have been strong enough to override or otherwise mask any potential anxiety triggered by the examiner's visual appearance. Additionally, without prior exposure to the examiner's speech, participants may have withheld assumptions about the examiner's linguistic profile. This early visual neutrality is further supported by the fact that, aside from P5, no participant spontaneously mentioned the examiner's appearance during the first half of the recall session, when the video from Test 1 was reviewed.

Second, in contrast to what was observed in the first test, results from the subsequent test revealed significantly higher pretest anxiety when participants were presented with the South Asian than the Caucasian examiner. Notably, the participants who demonstrated this effect had just completed the first test with the Caucasian examiner, so the change in their examiner's visual identity from Caucasian to Asian was now experienced in contrast rather than in isolation. In this context, the racialized expectations associated with the examiner's appearance may have been activated more strongly, particularly if the Caucasian guise functioned as a mental reference point for what an English-speaking test examiner "should" look like. One participant (P9) remarked that the second examiner "didn't look like the IELTS examiners" they had previously encountered or expected. Another participant (P22), whose pretest anxiety increased from 28 in Test 1 with the Caucasian examiner to 71 in Test 2 with the South Asian examiner, attributed this change to the examiner's "appearance." Similarly, P21, whose pretest anxiety rose from 30 with

the Caucasian examiner to 90 with the South Asian examiner, explained that this change in their anxiety was "just because I saw [the examiner] was an Indian."

The idea that a visual cue could elicit such a strong emotional response aligns with theories of RLS (Kang & Rubin, 2009), which suggest that visual characteristics—especially those that signal a person's racial or ethnic belonging—can shape language-related expectations before any speech is heard. These effects reflect a broader psychological tendency to interpret new information through the lens of pre-existing mental schemas (Langer & Abelson, 1974; Seligman et al., 1972). For instance, Darley and Gross (1983) showed that observers judged a child's academic ability differently based on the child's perceived socioeconomic status, even when evaluating the same behavior. In the present study, the appearance of the examiner may have similarly shaped participants' expectations about the difficulty or quality of the upcoming interaction, thereby increasing L2 speakers' anxiety prior to speaking. Several participants described experiencing anxiety upon seeing the South Asian examiner because they assumed she would be difficult to understand or might speak with a strong accent. One participant explained that they "naturally felt the anxiety" because they thought they might "get stuck" due to comprehension difficulties. Therefore, it seems that what Lippi-Green (2011, p. 94) called "imaginary accents" and "fictional communicative breakdowns" brought on by assumptions of an examiner's language ability could affect L2 speakers' emotional states at the pretest speaking stage—even before any speech input occurs.

Additionally, some participants associated the examiner's race not just with *how* they might speak, but with *what* they might ask. For P5, the image of the South Asian examiner prompted anxiety because they assumed she would ask "tricky questions" or ones focusing on unfamiliar cultural knowledge. As they described it: "I expected to encounter some tricky

questions… From the face of this examiner, I could tell that she would come up with some questions that I need to think about for a long while." They were concerned that the examiner might ask "questions related to their cultural background," which they felt unequipped to answer, hence the feeling of pretest anxiety. These associations between the examiner's presumed ethnicity and cultural knowledge and the test content further illustrate how and why speaking anxiety can be shaped by racialized expectations, even in a tightly controlled testing context.

Finally, there was a significant drop from pretest to posttest anxiety in Test 2, where those participants who demonstrated high anxiety before responding to the South Asian examiner generally felt less anxious after completing the speaking task. Interview data suggest that this decrease in anxiety was possibly driven by a mismatch between participants' initial expectations and their actual experience with the examiner's speech. Specifically, several participants assumed—based on the examiner's appearance—that she would speak with a strong accent or be difficult to understand. Once the speaking task began and participants heard clearly articulated English, this expectation was disconfirmed, leading to a reduction in posttest anxiety. For example, to explain why their anxiety dropped 58 points after responding to the South Asian examiner's prompts, P22 commented: "I thought that Indian [examiner] would speak Indian English, but she spoke quite clearly." Another participant, whose anxiety rating decreased dramatically before and after taking the second test with the South Asian examiner (90 vs. 35), explained this sharp decline as follows: "Because the examiner's accent is not very strong. There is no strong accent" (P21).

**RLS and Speakers' Perceptions of Examiners**

Although some participants showed evidence of RLS in their perception of the two examiners' speaking performances—such as through explicit comments about the South Asian

examiner's "Indian accent"—participants' assessments of the examiners' actual speech (e.g., in terms of accentedness and comprehensibility) and their willingness to be tested by either examiner did not reveal significant differences across the guises. Put differently, participants appeared to evaluate the speech itself objectively, with little influence from the examiner's appearance, and did not exhibit a systematic preference for one examiner over the other. This finding diverges from those reported in previous RLS studies where listeners' perceptions of speech were distorted by visual cues relating to race (e.g., Kang & Rubin, 2009; Kutlu et al., 2022; Rubin, 1992). Instead, when L2 speakers were given the chance to directly engage with the examiner's speech such as through listening and responding to the examiners' test prompts, those initial assumptions could be revised. In this case, listening—and especially close listening, given the context of the speaking test where participants naturally make a great effort to understand the target speech—allowed participants to recalibrate their expectations. This recalibration likely contributed not only to more accurate speech evaluations but also to reduced anxiety by the end of the test. As a result, participants converged in their perceptions of each examiner's speech and in their willingness to be assessed by either examiner, suggesting that meaningful engagement with speech may have overridden earlier stereotype-based expectations.

However, if any recalibration in participants' expectations took place, it was relatively short-lived. When participants were shown static images of the two examiners after completing both tests and asked to evaluate the nativeness of their English, a clear difference re-emerged: Participants considered the Caucasian examiner to speak more nativelike English than the South Asian examiner. In this sense, the examiner's image shown outside the interactive context continued to exert a residual influence on participants' linguistic judgment, despite their prior interactive experience that contradicted the stereotypical expectations triggered by the

examiner's appearance. Thus, while direct experience with an interlocutor's speech may reduce L2 speakers' reliance on stereotypes, this recalibration may not be durable once those auditory cues are no longer present. This pattern echoes findings in social psychology, which suggest that stereotypes, once activated, are not easily unlearned. Even when disconfirmed by direct experience, they often persist or re-emerge in later judgments (Darley & Gross, 1983; Devine, 1989; Nickerson, 1998). According to Devine (1989), stereotypes get activated rapidly and efficiently, so counter-stereotypical experiences often act less powerfully unless sustained, effortful corrections are made. In our case, any disconfirmatory evidence available from the examiner's actual speech was either less salient to participants or no longer easily accessible to enable them to effectively resist any stereotypes re-activated through visual cues to the examiner's racial or ethnic identity.

Lastly, participants reported generally similar levels of posttest anxiety, regardless of which examiner they encountered in either test. This suggests that any relationship between RLS and speaking anxiety was confined to the anticipatory stage—prior to the onset of speech—when participants had not yet interacted with the examiner and thus had no experience with her speech. One possible explanation is that anxiety tends to peak before speaking begins, when L2 speakers face the greatest sense of evaluative pressure and social judgment (Horwitz et al., 1986). In explaining their pretest anxiety with the Caucasian examiner, for example, P5 illustrated how the imaginary negative judgement could create emotional tension: "I think she might be a native speaker of English… I feel a bit nervous… Because some of my errors may not be a problem, but for those native speakers it might be a glaring mistake." In this pre-speaking phase, visual cues to the examiner's identity seemed to have exerted a disproportionate influence on P5's

emotional state, especially when combined with the limited contextual information they felt they had about the test or their performance at that time.

Prior research using moment-by-moment tracking methods has shown that speaking anxiety often decreases when speakers begin the speaking task and shift their cognitive focus from fear of evaluation to message delivery (Gregersen et al., 2014; Young, 1986). In the present study, it is likely that once participants began responding, the influence of visually-cued stereotypes diminished as participants became cognitively and linguistically engaged. Additionally, the structured format of the task—with noninteractive prompts and predictable timing—may have reduced opportunities for stereotype-based discomfort to escalate during the speech itself. That said, this conclusion is limited by the fact that anxiety was only self-rated before and after the speaking task in each test. It remains possible that stereotype-driven pressure continued to affect participants during their speech production—perhaps during at least some of their responses to the three consecutive prompts in each test—in ways not captured through self-rated assessments. Future research might explore this question more directly by analyzing real-time speech content (e.g., hesitation patterns, lexical choice, fluency, etc.) or by tracking anxiety moment-to-moment through idiodynamic or physiological methods.

## Implications for L2 Learning and Assessment

The findings in this study suggest that RLS, which refers to people's perceptions about a speaker's linguistic performance based on the visual cues to the speaker's racial or ethnic identity, can contribute to L2 speakers' anxiety, most likely at the anticipatory stage. This effect appears most salient when L2 speakers' expectations of what an English test examiner "should" look like are disrupted, for instance, after the more stereotype-congruent Caucasian examiner in the initial test was replaced by the less stereotype-congruent South Asian examiner. Based on these findings, two implications emerge for the broader contexts of L2 learning and assessment. First, the findings suggest that L2 speakers should be supported in recognizing and managing the implicit biases they might bring to speaking tasks and how these biases could possibly heighten their feelings of anxiety before speaking. In fact, several participants in this study demonstrated some awareness of their assumptions during the recall session, acknowledging the role of their biases and stereotypes, gained through the media or personal experience, in expressing judgments about the examiners. However, this awareness alone did not seem sufficient to neutralize their emotional response. For instance, one participant who, despite admitting that the South Asian examiner actually spoke clearly, which contradicted their initial assumption, still expressed a preference for a "White" examiner because this examiner identity appears congruent with the expected nativelikeness of their speech. What emerges from these comments is the entrenched nature of L2 speaker biases (Darley & Gross, 1983; Devine, 1989; Kang & Rubin, 2009). Language instruction should therefore not only help L2 speakers become aware of bias but also offer concrete strategies for managing it. These could include increased exposure to racially diverse English speakers (either in real life interactions or through the media), reflective discussions on language and identity, and guided activities unpacking and problematizing what

"standard" or "nativelike" speech really entails (Higgins, 2003; Lippi-Green, 2011). By addressing RLS at both the cognitive and experiential levels, educators can help learners build resilience against stereotype-based anxiety and improve their confidence in diverse communicative contexts.

Second, the findings point to a clear need to normalize racial diversity among language examiners in both pedagogical and assessment contexts. In fact, such a call has already been well documented in the language teaching profession (e.g., Amin, 1997; Golombek & Jordan, 2005; Kubota & Lin, 2006; Ruecker & Ives, 2015), where "English and Whiteness are thornily intertwined," most likely because "the spread of the English language across the globe was historically connected to the international political power of White people" (Motha, 2006, p. 496). Although few studies to date have focused specifically on whether this phenomenon extends to L2 assessment contexts (e.g., Hutabarat, 2023), several participants in this study expressed a preference for White examiners in oral exams, citing this as the "norm" they had come to expect. For example, when describing the South Asian examiner during the recall session, one participant remarked: "She's not like the first examiner, who... How do I put this? She doesn't look like the IELTS examiners I've had." For another participant, this belief was not grounded in personal experience but shaped by online narratives: "Because I read a lot of news and posts about Indian examiners, and their accent being a problem… a lot of people complain about the Indian accents, this and that about the examiner." When commenting on the Caucasian examiner, the same participant noted: "The way this examiner looks, [she] is a very typical examiner for TOEFL or IELTS." Clearly, for these participants, a non-White examiner did not conform visually to what they considered a "legitimate" English-speaking test examiner.

Considering that these (often subconscious) preferences can generate pre-speaking anxiety, it is critical to challenge the default mental association between Whiteness and linguistic authority (Lippi-Green, 2011; Motha, 2006). One way to address this is by intentionally increasing racial and ethnic diversity among examiners in standardized English tests such as IELTS or TOEFL, in both in-person and remote formats. Regular exposure to diverse examiner identities may help recalibrate test-takers' expectations and ideally reduce their expectation-triggered anxiety. Additionally, incorporating a brief, non-evaluative interaction between the examiner and the test-taker before the formal task begins—such as through a short greeting or a warm-up exchange—may help mitigate anxiety by allowing L2 speakers to experience the examiner's actual speech. As revealed through the findings of this study, L2 speakers' often negative assumptions about an examiner's language ability could be addressed after engaging with the examiner's actual speech. In fact, in a study focusing on moment-to-moment fluctuations in L2 speakers' emotions, Shirvan and Talebzadeh's (2017) showed that even a brief interpersonal exchange that lasts just a few minutes can lead to a noticeable reduction in L2 speakers' anxiety. Therefore, including a non-evaluative pretest interaction between examiners and test-takers could be a practical and low-cost solution to disconfirm visually-triggered assumptions and ease anticipatory speaking anxiety. Together, these structural and procedural adjustments may help reduce the emotional burden associated with speaking assessments.

## Limitations and Future Work

As an initial, exploratory study that extends research on RLS beyond L2 speakers' comprehension, the current work has several limitations and thus calls for further research. One key limitation concerns the measurement of speaking anxiety. In this study, anxiety was measured only twice—once before and once after each speaking task—using participant self-reports. While this approach captured broad changes in participants' emotional states, it likely missed the full range of fluctuations and dynamics that can occur during speech production. More fine-grained methods, such as moment-to-moment self-ratings or physiological tracking could reveal how anxiety evolves across different stages of interaction (Gregersen et al., 2014; Lindberg et al., 2023). Alternatively, a video recording of participants' behavioral reactions might also provide a deeper understanding of their speaking anxiety. As documented previously (e.g., Gregersen, 2005; Lindberg et al., 2021), anxious L2 speakers behave differently at the nonverbal level, for example, revealing their enhanced anxiety states through a more rigid posture, more glancing away and blinking, less smiling, and more self-adaptive behaviors such as touching their hair or face and manipulating objects.

Furthermore, although subjective anxiety ratings offer insight into test-takers' internal experiences, they may not fully align with observable linguistic indicators of anxiety. Although the current study included a preliminary analysis of participants' response quantity, with no significant differences detected in amount of speech across the two examiner guises (see Appendix G for full descriptive data), future research could examine speech output more closely. For example, independent raters could evaluate various aspects of L2 speakers' production such as lexical richness, syntactic complexity, fluency, and pronunciation to determine whether anxiety subtly influences performance quality in ways not captured by self-report alone (Pérez

Castillejo, 2019; Sosa-López & Mora, 2022). According to MacIntyre and Gardner (1994), speaking anxiety is negatively correlated with language performance; when performing a task, anxious speakers could demonstrate a smaller vocabulary size and find it harder to retrieve appropriate words. Therefore, a close look into L2 speakers' output quality may generate different insights into the role of RLS in speaking anxiety.

One last point about the measurement of anxiety is that it might be useful for future research to account for individual differences in baseline anxiety levels. Some participants in this study tended to be more anxious than others, presumably because anxiety was part of their relatively stable personality trait (Spielberger, 1966). For example, one participant found it difficult to even "express themselves" during a speaking test and may "experience emotional breakdowns or cry" afterward (P14). The experiences of these naturally high-anxious L2 speakers may therefore differ significantly from those who are not inherently as anxious about speaking or test-taking. According to prior research, speakers who are more test-anxious tend to give more attention to self-preoccupied worry and task cues in evaluative situations, whereas less test-anxious speakers may focus more on the task themselves (Young, 1986). Therefore, future research is encouraged to incorporate an established anxiety scale such as the FLCAS developed by Horwitz et al. (1986) to assess the trait anxiety of participants aside from measuring their state anxiety during the test. Including such measures would allow future studies to differentiate between situation-specific anxiety (e.g., triggered by an examiner's appearance) and more stable individual predispositions, leading to a more nuanced understanding of how RLS is associated with speakers' affective states.

Another aspect that future research could continue to refine is the format of the speaking tasks. Although presenting the task as a mock English speaking test aligns well with the goal of

this study—given that formal testing contexts are known to provoke speaking anxiety (Horwitz et al., 1986)—greater variation in task type could offer additional insight into how RLS operates in different communicative settings. For example, future studies could incorporate alternative task formats that maintain the presence of racially diverse individuals while reducing the evaluative intensity of the task itself. For instance, picture description tasks or storytelling could be adopted to prompt participants' oral responses. These tasks, which require more interpretation and imagination, could in fact be more reflective of speaking anxiety (Steinberg & Horwitz, 1986). Lastly, future work may explore speakers' anxiety in interactive tasks such as conversational exchanges between test-takers and the examiner. Interpersonal dynamics are known to shape language anxiety (Mak, 2011; Shirvan & Talebzadeh, 2017; Woodrow, 2006), and it remains an open question whether RLS effects on speakers' anxiety and their performance would be amplified, diminished, or altered in real-time exchanges where examiner responsiveness, tone, and nonverbal cues come into play. The lack of real-time communication in this study, therefore, may have missed interactional signs of anxiety or subtle shifts in participant behavior in response to the examiner's perceived identity. One participant, who found themselves nervous answering one question because they did not know what to say, expressed that they may have felt different if they were given the chance to ask the examiner to "narrow down the question." Knowing this, we expect that RLS effects on anxiety—such as through accommodation strategies or changes in engagement—might emerge differently in live conversations.

## Conclusion

Motivated by RLS research, including Rubin's (1992) influential study, the current work aimed to extend the scope of inquiry by examining how L2 speakers' assumptions about test examiners—based on the visual cues to their racial identity—might influence their speaking anxiety. While this exploratory study does not allow for definitive conclusions about the causal relationship between RLS and speaking anxiety, it did reveal a consistent pattern: seeing a non-White examiner, particularly one who visually contrasted with perceived or expected norms, tended to elevate L2 speakers' anxiety prior to speaking. The study therefore underscores the powerful role that visual cues—and the social, language-related assumptions they activate—can play in shaping emotional responses among L2 speakers. In the absence of test examiners' actual speech, their racialized appearance was sufficient to provoke discomfort before L2 speakers began speaking, revealing how deeply internalized ideologies about language and identity could shape not only their perceptions but also their affective response before performing an evaluative task. By bringing attention to this new dimension of RLS, the study highlights the need for L2 speakers to confront their stereotyped beliefs—because stereotypes are often too light to be noticed and, if left unquestioned, eventually become too heavy to be broken.

## References


Abrams, D., & Hogg, M. A. (1987). Language attitudes, frames of reference, and social identity: A Scottish dimension. *Journal of Language and Social psychology*, *6*(3-4), 201-213. https://doi.org/10.1177/0261927X8763004

Amin, N. (1997). Race and the identity of the nonnative ESL teacher. *TESOL quarterly*, *31*(3), 580-583. https://doi.org/10.2307/3587841

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

Baratta, A. (2017). Accent and linguistic prejudice within British teacher training. *Journal of Language, Identity & Education*, *16*(6), 416-423. https://doi.org/10.1080/15348458.2017.1359608

Bielak, J. (2022). To what extent are foreign language anxiety and foreign language enjoyment related to L2 fluency? An investigation of task-specific emotions and breakdown and speed fluency in an oral task. *Language Teaching Research*, *00* (0) 1-30. https://doi.org/10.1177/13621688221079319

Burgers, C., & Beukeboom, C. J. (2020). How language contributes to stereotype formation: Combined effects of label types and negation use in behavior descriptions. *Journal of Language and Social Psychology*, *39*(4), 438-456. https://doi.org/10.1177/0261927X20933320

Campbell-Kibler, K. (2005). *Listener perceptions of sociolinguistic variables: The case of (ING)*. [Doctoral dissertation, Stanford University]. ProQuest Dissertations Publishing.

Chun, S. Y. (2014). EFL learners' beliefs about native and non-native English-speaking teachers: perceived strengths, weaknesses, and preferences. *Journal of Multilingual and*

*Multicultural Development*, *35*(6), 563-579. https://doi.org/10.1080/01434632.2014.889141

Clark, E., & Paran, A. (2007). The employability of non-native-speaker teachers of EFL: A UK survey. *System*, *35*(4), 407-430. https://doi.org/10.1016/j.system.2007.05.002

Cohen, Y., & Norst, M. J. (1989). Fear, dependence and loss of self-esteem: Affective barriers in second language learning among adults. *RELC Journal*, *20*(2), 61-77. https://doi.org/10.1177/003368828902000206

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly*, *33*(2), 185-209. https://doi.org/10.2307/3587717

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20. https://doi.org/10.1037/0022-3514.44.1.20

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5. https://doi.org/10.1037//0022-3514.56.1.5

Dewaele, J. M., Petrides, K. V., & Furnham, A. (2008). Effects of trait emotional intelligence and socio-biographical variables on communicative anxiety and foreign language anxiety among adult multilinguals: A review and empirical investigation. *Language Learning*, *58*(4), 911-960. https://doi.org/10.1111/j.1467-9922.2008.00482.x

Dragojevic, M., & Goatley-Soan, S. (2022). Americans' attitudes toward foreign accents: Evaluative hierarchies and underlying processes. *Journal of Multilingual and Multicultural Development*, *43*(2), 167-181. https://doi.org/10.1080/01434632.2020.1735402

Edwards, J. (1999). Refining our understanding of language attitudes. *Journal of Language and Social Psychology*, *18*(1), 101-110. https://doi.org/10.1177/0261927X9901800100

Galloway, N. (2014). "I get paid for my American accent": The story of one multilingual English teacher (MET) in Japan. *Englishes in Practice*, *1*(1), 1-30. https://doi.org/10.2478/eip-2014-0001

Giles, H., Bourhis, R., & Davies, A. (1979). Prestige speech styles: the imposed norm and inherent value hypotheses. *Language and Society: Anthropological Issues*, 589-596. https://doi.org/10.1515/9783110806489.589

Giles, H., Bourhis, R., Lewis, A., & Trudgill, P. (1974). The imposed norm hypothesis: A validation. *Quarterly Journal of Speech*, *60*(4), 405-410. https://doi.org/10.1080/00335637409383249

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

Golombek, P., & Jordan, S. R. (2005). Becoming "black lambs" not "parrots": A poststructuralist orientation to intelligibility and identity. *TESOL Quarterly*, *39*(3), 513-533. https://doi.org/10.2307/3588492

Gregersen, T. S. (2005). Nonverbal cues: Clues to the detection of foreign language anxiety. *Foreign Language Annals*, *38*(3), 388-400.

Gregersen, T., MacIntyre, P. D., & Meza, M. D. (2014). The motion of emotion: Idiodynamic case studies of learners' foreign language anxiety. *The Modern Language Journal*, *98*(2), 574-588. https://doi.org/10.1111/modl.12084

Hansen Edwards, J. G., Zampini, M. L., & Cunningham, C. (2018). The accentedness, comprehensibility, and intelligibility of Asian Englishes. *World Englishes*, *37*(4), 538-557. https://doi.org/10.1111/weng.12344

Hashemi, M. (2011). Language stress and anxiety among the English language learners. *Procedia – Social and Behavioral Sciences*, *30*, 1811–1816. https://doi.org/10.1016/j.sbspro.2011.10.349

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

Higgins, C. (2003). "Ownership" of English in the Outer Circle: An alternative to the NS-NNS dichotomy. *TESOL Quarterly*, *37*(4), 615-644. https://doi.org/10.2307/3588215

Hilton, N. H., Gooskens, C., Schüppert, A., & Tang, C. (2022). Is Swedish more beautiful than Danish? Matched guise investigations with unknown languages. *Nordic Journal of Linguistics*, *45*(1), 30-48. https://doi.org/10.1017/S0332586521000068

Horwitz, E. K. (2010). Foreign and second language anxiety. *Language Teaching*, *43*(2), 154-167. https://doi.org/10.1017/S026144480999036X

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, *70*(2), 125-132. https://doi.org/10.2307/327317

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

Hu, G., & Su, J. (2015). The effect of native/non-native information on non-native listeners' comprehension. *Language Awareness*, *24*(3), 273-281. http://dx.doi.org/10.1080/09658416.2015.1077853

Hutabarat, P. (2023). Becoming IELTS Examiners: Demystifying Native-Speakerism in the Area of English Language Testing. *Journal of Education and Teaching (JET), 4*(1), 50-68.

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., Yaw, K., & Kostromitina, M. (2023). The effects of situational contexts and occupational roles on listeners' judgements on accented speech. *Psychology of Language and Communication*, *27*(1), 1-22. https://doi.org/10.58734/plc-2023-0001

Kormos, J., & Préfontaine, Y. (2017). Affective factors influencing fluent performance: French learners' appraisals of second language speech tasks. *Language Teaching Research*, *21*(6), 699-716. https://doi.org/10.1177/1362168816683562

Kubota, R., & Lin, A. (2006). Race and TESOL: Introduction to concepts and theories. *TESOL Quarterly*, *40*(3), 471-493. https://doi.org/10.2307/40264540

Kutlu, E., Tiv, M., Wulff, S., & Titone, D. (2022). The impact of race on speech perception and accentedness judgments in racially diverse and non-diverse groups. *Applied Linguistics*, *43*(5), 867-890. https://doi.org/10.1093/applin/amab072

Ladegaard, H. J. (1998). National stereotypes and language attitudes: The perception of British, American and Australian language and culture in Denmark. *Language & Communication*, *18*(4), 251-274. https://doi.org/10.1016/S0271-5309(98)00008-1

Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology*, *60*(1), 44.

Langer, E. J., & Abelson, R. P. (1974). A patient by any other name...: Clinician group difference in labeling bias. *Journal of Consulting and Clinical Psychology*, *42*(1), 4.

Lasagabaster, D., & Sierra, J. M. (2002). University students' perceptions of native and non-native speaker teachers of English. *Language Awareness*, *11*(2), 132-142. https://doi.org/10.1080/09658410208667051

Lee, B. J., & Bailey, J. L. (2023). Assumptions of speaker ethnicity and the effect on ratings of accentedness, comprehensibility, and intelligibility. *Language Awareness*, *32*(2), 301-322. https://doi.org/10.1080/09658416.2022.2091143

Lee, T. L., & Fiske, S. T. (2006). Not an outgroup, not yet an ingroup: Immigrants in the stereotype content model. *International Journal of Intercultural Relations*, *30*(6), 751-768. https://doi.org/10.1016/j.ijintrel.2006.06.005

Levis, J. M., Sonsaat, S., Link, S., & Barriuso, T. A. (2016). Native and nonnative teachers of L2 pronunciation: Effects on learner performance. *TESOL Quarterly*, *50*(4), 894-931. https://doi.org/10.1002/tesq.272

Lindberg, R., McDonough, K., & Trofimovich, P. (2021). Investigating verbal and nonverbal indicators of physiological response during second language interaction. *Applied Psycholinguistics*, *42*(6), 1403-1425. https://doi.org/10.1017/S014271642100028X

Lindberg, R., McDonough, K., & Trofimovich, P. (2023). Second language anxiety in conversation and its relationship with speakers' perceptions of the interaction and their social networks. *Studies in Second Language Acquisition*, *45*(5), 1413-1426. https://doi.org/10.1017/S0272263122000523

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. (2011). *English with an accent: language, ideology and discrimination in the United States* (Second edition). Routledge.

Liu, M. (2006). Anxiety in Chinese EFL students at different proficiency levels. *System*, *34*(3), 301-316. https://doi.org/10.1016/j.system.2006.04.004

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, *47*, 265–287. https://doi.org/10.1111/0023-8333.81997008

Mak, B. (2011). An exploration of speaking-in-class anxiety with Chinese ESL learners. *System*, *39*(2), 202-214. https://doi.org/10.1016/j.system.2011.04.002

McDonough, K., Trofimovich, P., Tekin, O., & Sato, M. (2022). Exploring linguistic stereotyping of international students at a Canadian university. *Journal of Multilingual and Multicultural Development*, 1-16. https://doi.org/10.1080/01434632.2022.2115049

McKenzie, R. M. (2008). The role of variety recognition in Japanese university students' attitudes towards English speech varieties. *Journal of Multilingual and Multicultural Development*, *29*(2), 139-153. https://doi.org/10.2167/jmmd565.0

Motha, S. (2006). Racializing ESOL teacher identities in US K-12 public schools. *TESOL Quarterly*, 495-518. https://doi.org/10.2307/40264541

Nagle, C., Trofimovich, P., & Bergeron, A. (2019). Toward a dynamic view of second language comprehensibility. *Studies in Second Language Acquisition*, *41*(4), 647-672. https://doi.org/10.1017/S0272263119000044

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175-220. https://doi.org/10.1037/1089-2680.2.2.175

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

Peirce, B. N. (1995). Social identity, investment, and language learning. *TESOL quarterly*, *29*(1), 9-31. https://doi.org/10.2307/3587803

Pérez Castillejo, S. (2019). The role of foreign language anxiety on L2 utterance fluency during a final exam. *Language Testing*, *36*(3), 327-345. https://doi.org/10.1177/0265532218777783

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878-912. https://doi.org/10.1111/lang.12079

Rampton, B. (1990). Displacing the "native speaker": Expertise, affiliation and inheritance. *ELT Journal*, *44*, 97–101. https://doi.org/10.1093/eltj/44.2.97

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135-147. https://doi.org/10.1016/j.edurev.2010.12.001

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

Ruecker, T., & Ives, L. (2015). White native English speakers needed: The rhetorical construction of privilege in online teacher recruitment spaces. *TESOL Quarterly*, *49*(4), 733-756. https://doi.org/10.1002/tesq.195

Ryan, E. B. (1983). Social psychological mechanisms underlying native speaker evaluations of non-native speech. *Studies in Second Language Acquisition*, *5*(2), 148-159.

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*(2), 217-240. https://doi.org/10.1017/S0142716414000502

Sandelowski, M. (2000). Whatever happened to qualitative description?. *Research in Nursing & Health*, *23*(4), 334-340. https://doi.org/10.1002/1098-240X(200008)23:4<334::AID-NUR9>3.0.CO;2-G

Schüppert, A., Hilton, N. H., & Gooskens, C. (2015). Swedish is beautiful, Danish is ugly? Investigating the link between language attitudes and spoken word recognition. *Linguistics*, *53*(2), 375-403. https://doi.org/10.1515/ling-2015-0003

Scovel, T. (1978). The effect of affect on foreign language learning: A review of the anxiety research. *Language Learning*, *28*(1), 129-142.

Seligman, C. R., Tucker, G. R., & Lambert, W. E. (1972). The effects of speech style and other attributes on teachers' attitudes toward pupils. *Language in Society*, *1*(1), 131-142.

Shirvan, M. E., & Talebzadeh, N. (2017). English as a foreign language learners' anxiety and interlocutors' status and familiarity: An idiodynamic perspective. *Polish Psychological Bulletin*, *48*, 489–503. https://doi.org/10.1515/ppb-2017-0056

Simon-Maeda, A. (2004). The complex construction of professional identities: Female EFL educators in Japan speak out. *TESOL Quarterly*, *38*, 405–436. https://doi.org/10.2307/3588347

Sosa-López, G., & Mora, J. C. (2022). The Role of Speaking Anxiety on L2 English Speaking Fluency, Accuracy and Complexity. *Pronunciation in Second Language Learning and Teaching Proceedings*, *12*(1). https://doi.org/10.31274/psllt.13362

Spielberger, C. D., (1966). Theory and Research on Anxiety. In C. D. Spielberger (Eds.), *Anxiety and Behavior* (pp. 3-20). Academic Press.

Spielberger, C. D., Gonzalez-Reigosa, F., Martinez-Urrutia, A., Natalicio, L., & Natalicio, D. S. (1971). Development of the Spanish edition of the state-trait anxiety inventory. *Interamerican Journal of Psychology*, *5*(3-4), 145-158.

Steinberg, F. S., & Horwitz, E. K. (1986). The effect of induced anxiety on the denotative and interpretive content of second language speech. *TESOL Quarterly*, *20*(1), 131-136. https://doi.org/10.2307/3586395

Tekin, O., & Trofimovich, P. (2023). En français or in English? Examining perceived social roles of international students in response to their French and English speech. The *Canadian Modern Language Review*, *79*(3), 204-227. https://doi.org/10.3138/cmlr-2022-0037

The Jamovi Project (2025). jamovi (Version 2.6) [Computer Software]. Retrieved from https://www.jamovi.org.

Thorne, S., Kirkham, S. R., & MacDonald-Emes, J. (1997). Interpretive description: a noncategorical qualitative alternative for developing nursing knowledge. *Research in Nursing & Health*, *20*(2), 169-177. https://doi.org/10.1002/(SICI)1098-240X(199704)20:2<169::AID-NUR9>3.0.CO;2-I

Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2023). Frequency of exposure influences accentedness and comprehensibility in learners' pronunciation of second language words. *Language Learning*, *73*(1), 84-125. https://doi.org/10.1111/lang.12517

Woodrow, L. (2006). Anxiety and speaking English as a second language. *RELC Journal*, *37*(3), 308-328. https://doi.org/10.1177/003368820607131

Young, D. J. (1986). The relationship between anxiety and foreign language oral proficiency ratings. *Foreign Language Annals*, *19*(5), 439-445. https://doi.org/10.1111/j.1944-9720.1986.tb01032.x

# Appendixes

## Appendix A

### Background Questionnaire

**General Information**:

1. Full Name: _____
2. Age: _____
3. Gender: _____
4. Country of Origin: _____
5. Birthplace (city, province): _____
6. When did you arrive in Montreal (month, year)? _____
7. In what term did you start your studies at Concordia (e.g., Fall 2024)? _____
8. What program are you in (e.g., MA in Journalism)? _____

**Language and Language Experience**:

1. What is your first language (mother tongue)? _____
2. Is English your second language? Yes No
3. What other language (s) do you know? Please indicate your proficiency level (e.g., Intermediate French): _____
4. How long have you been learning English? _____ years
5. Among the following standardized English tests, circle the one you have taken (if you have taken more than one test, please indicate the most recent one).
   a. Academic IELTS
   b. TOEFL
   c. Duolingo English Test
   d. Others: _____
6. When did you take the test (month, year)? _____
7. What was your overall score? _____
8. What was your speaking score? _____
9. Before coming to Canada, have you ever studied or lived in other English-speaking countries? Yes No
   If yes, please indicate the country and the length of stay (e.g., the U.S., 6 months):
   _____

10. Please self-rate your **current** level of English proficiency in each of the following areas by **circling** the appropriate number:

| | *Beginner* | | | | *Intermediate* | | | *Nativelike* | |
|---|---|---|---|---|---|---|---|---|---|
| **Speaking** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| **Listening** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| **Reading** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| **Writing** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

**Appendix B**

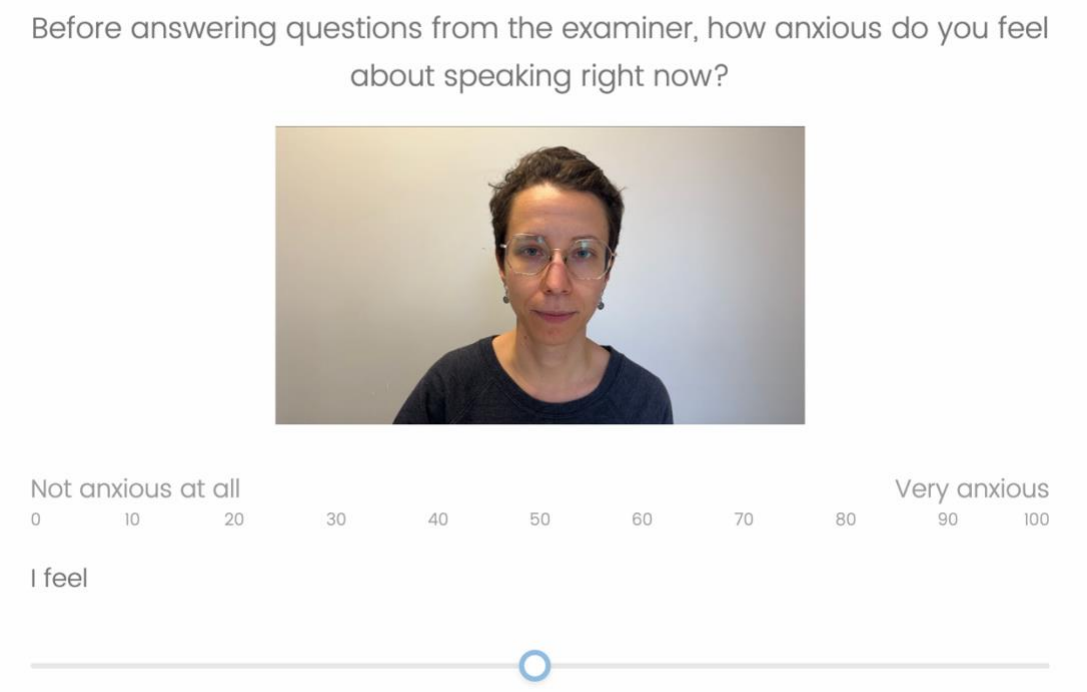
## Prompt Questions

| | AI | Social media | Making friends (Practice) |
|---|---|---|---|
| Opening | Let's talk about artificial intelligence, or AI. | Let's talk about social media. | Let's talk about making friends. |
| Q1 | What are the advantages and disadvantages of using AI in university education? | What are the positive and negative sides of limiting access to social media for young people? | What are the advantages and disadvantages of having friends who are similar to you? |
| Q2 | What are the ways in which university students usually use AI? Can you give specific examples? | What do young adults normally do on social media? Can you give specific examples? | / |
| Q3 | Do you think AI will have an influence on human creativity? Why or why not? | Do you think social media will change human relationships? Why or why not? | / |

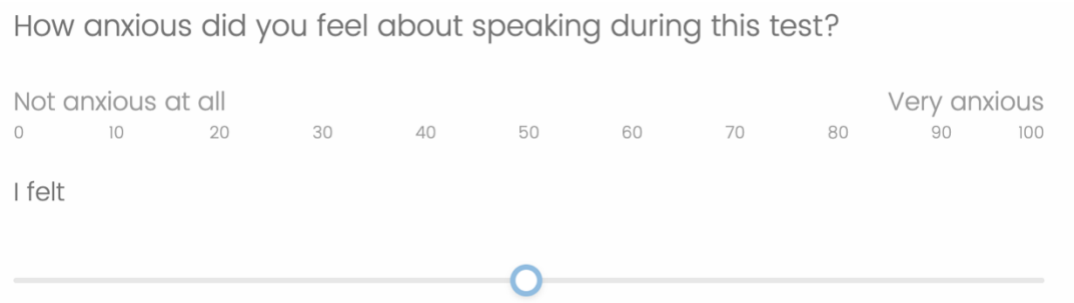
**Appendix C**

## Rating Scales

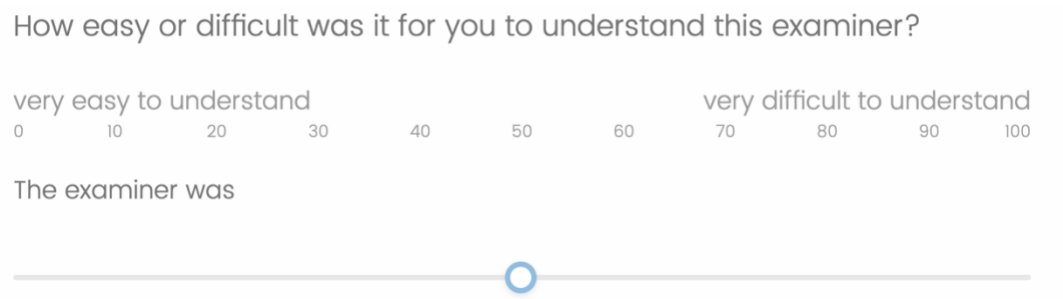
Pretest speaking anxiety:



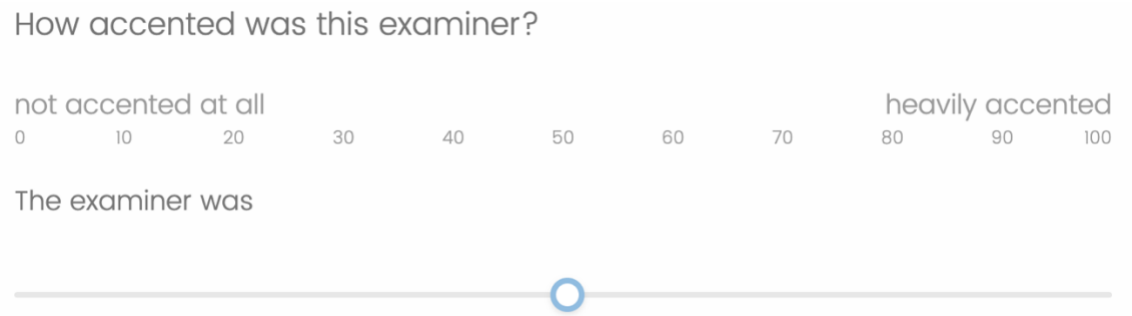

Posttest speaking anxiety:



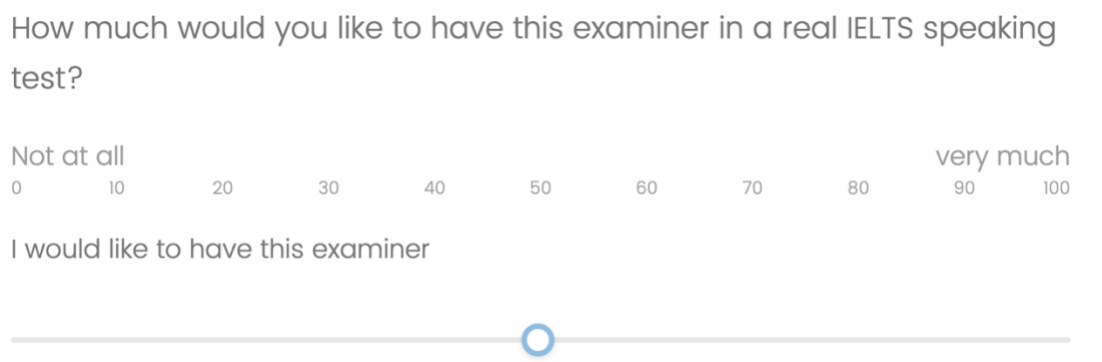

Examiner's comprehensibility:



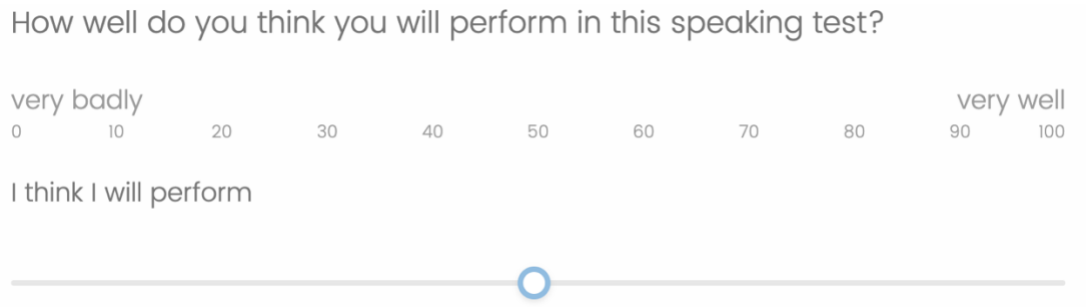

Examiner's accentedness:



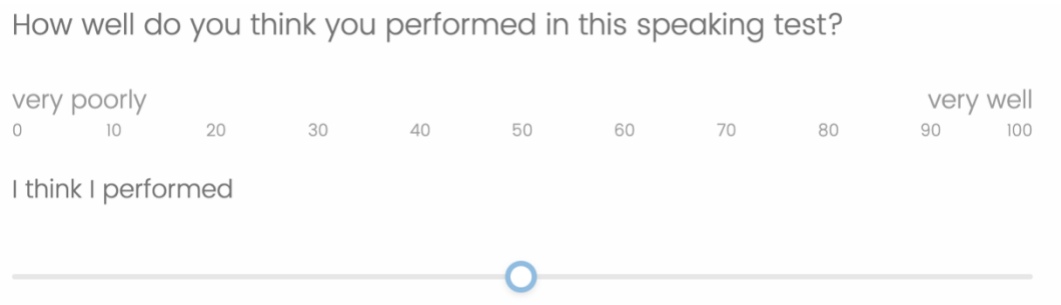

Willingness to have the examiner in a real exam:



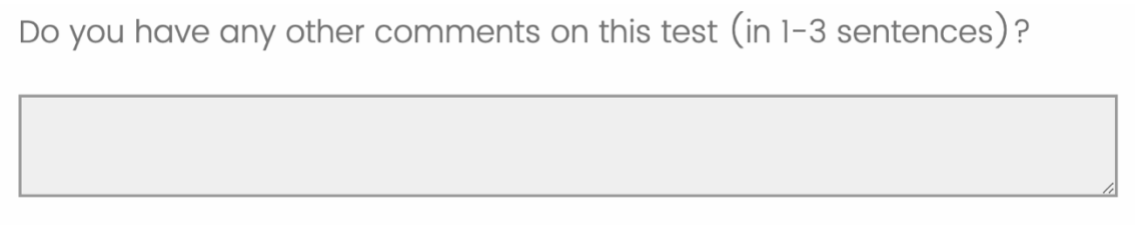

Anticipation of speaking performance before test:




Self-assessment of speaking performance after test:




Comments on the test:



**Appendix D**

**Post-test Survey**

1. Racial categorization of the examiners:

Which of these descriptors would you choose to describe this examiner?

- ○ East Asian (Chinese, Japanese, Korean, Mongolian, etc.)
- ○ Southeast Asian (Cambodian, Indonesian, Thai, Vietnamese, etc.)
- ○ South Asian (Indian, Tamil, Bangladeshi, Bengali, Punjabi, etc.)
- ○ Latin / Central / South American (Peruvian, Colombian, Argentinian, Mexican, etc.)
- ○ Middle Eastern (Iraqi, Syrian, Jordanian, Saudi Arabian, Lebanese, etc.)
- ○ European (Irish, Dutch, Norwegian, Russian, etc.)
- ○ African (Nigerian, Kenyan, Zambian, etc.)

Do you think this examiner is a native speaker of English?

second language speaker of English native speaker of English

I think the examiner is a...

Which of these descriptors would you choose to describe this examiner?

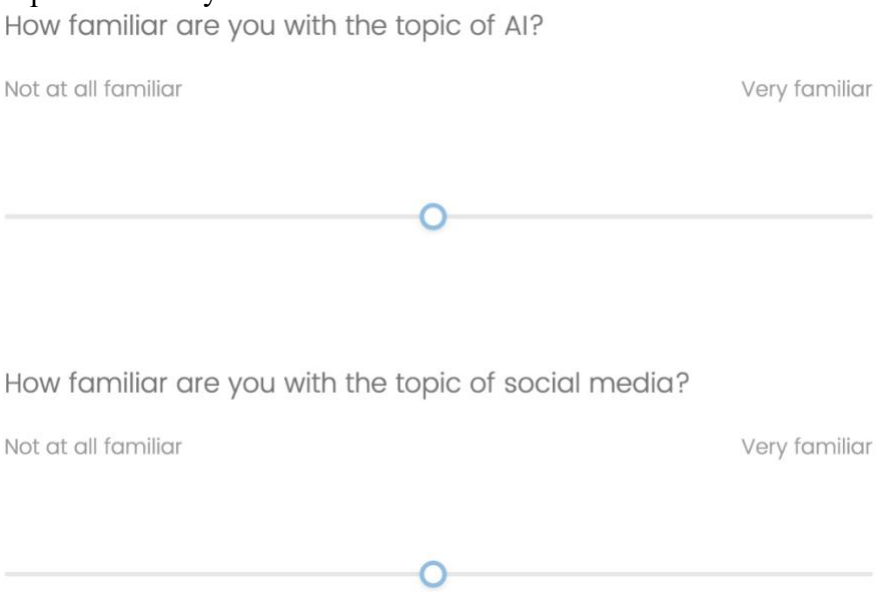
- ◯ East Asian (Chinese, Japanese, Korean, Mongolian, etc.)
- ◯ Southeast Asian (Cambodian, Indonesian, Thai, Vietnamese, etc.)
- ◯ South Asian (Indian, Pakistan, Bangladeshi, Bengali, Punjabi, etc.)
- ◯ Latin / Central / South American (Peruvian, Colombian, Argentinian, Mexican, etc.)
- ◯ Middle Eastern (Iraqi, Syrian, Jordanian, Saudi Arabian, Lebanese, etc.)
- ◯ European (Irish, Dutch, Norwegian, Russian, etc.)
- ◯ African (Nigerian, Kenyan, Zambian, etc.)

Do you think this examiner is a native speaker of English?

second language speaker of English native speaker of English

I think the examiner is a...

2. Topic Familiarity



**Appendix E**

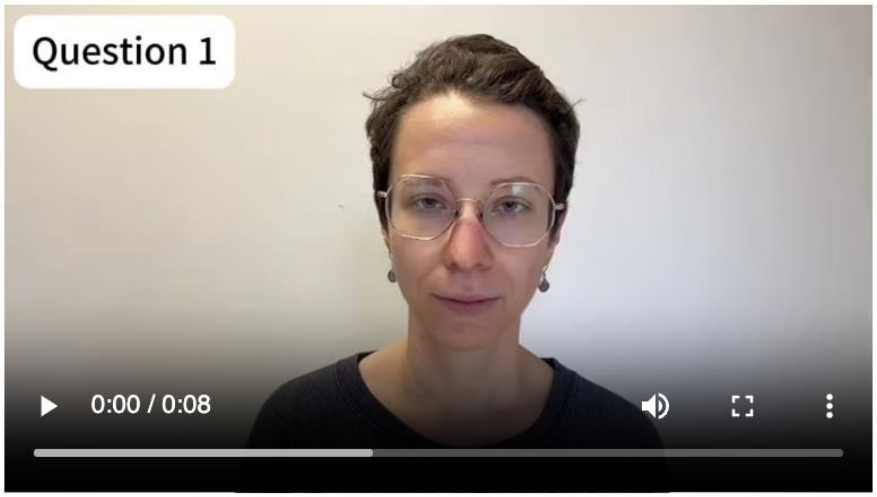
**Speaking Test Interface**

Please play the video, listen to the question, and start your response after the beep.

**The video can only be played once**, so please pay close attention.

**You will have 45 seconds to speak.**

Please **speak as long as possible** until you are stopped.



**Appendix F**

**Interview Questions (English)**

**Before** play the video for recall:

Overall, how did you feel during the two speaking tests?

- Were there specific moments or challenges during that test that made you more anxious?

**During** video-aided recall:

- Your anxiety before the test was X, could you explain why it was so high/low?
    - **(If mention examiner) What was it about the examiner that made you anxious?**
    - **(If mention examiner) Did you have experience interacting with a similar examiner?**
- Your self-rated performance before the test was X, could you explain why you thought you would/ would not perform well?
- Your anxiety after this test was X, could you explain why it was so high/low?
- Your self-rated performance after the test was X, could you explain why you thought you performed well/ did not perform well?
    - **Do you think your performance is affected by your high/ low anxiety?**
- Why did you think this examiner is easy/difficult to understand?
    - **Did that make you feel more anxious during the exam?**
- Why did you feel this examiner is heavily accented?
    - **Did that make you feel more anxious during the exam?**
- Why do you want/not want this examiner in a real English-speaking test?
    - **(If mention examiner's race) Did you have good/bad experience interacting with examiner of the same race?**

**Lastly:** "Your anxiety before the first test was X, and it was Y before the second test. Can you explain why there is a relatively big gap?"

**Interview Questions (Chinese)**

播放回顾视频**前**：
总体来说，你在这两次口语测试中感觉如何？

- 在测试过程中，是否有特定的时刻或挑战让你感到更焦虑？

在观看视频并进行**回顾时**：

- 你在测试前的焦虑值是 X，能解释一下为什么这么高/低吗?
  - **（如果提到考官）关于这位考官，是什么让你感到焦虑？**
  - **（如果提到考官）你是否有过与类似考官互动的经历？**
- 你在测试前对自己表现的评分是 X，能解释一下你为什么觉得自己会表现好 / 不好吗？
- 这场测试之后你的焦虑值是 X，能说说为什么这么高/低吗？
- 这场测试之后你对自己表现的评分是 X，能解释一下你为什么觉得自己表现好 / 不好吗？
  - 你觉得你的表现是否受到你焦虑程度的影响？
- 你为什么觉得这位考官容易 / 难以理解？
  - 这是否让你在考试过程中感到更焦虑？
- 你为什么觉得这位考官口音很重？
  - 这是否让你在考试过程中感到更焦虑？
- 你为什么想 / 不想让这位考官出现在真正的英语口语测试中？
  - （如果提到考官种族）你是否曾与相同种族的考官有过好的 / 不好的互动经历？

**最后一个问题：**

你的焦虑值在第一个测试前是 X，在第二个测试前是 Y。你能解释一下为什么这两个数值差距比较大吗？

## Appendix G

### Descriptive Statistics for Response Length

Descriptive Statistics for Response in Test 1

| | T1_examiner | N | Mean | SD |
|---|---|---|---|---|
| **T1Q1** | **Caucasian** | 20 | 84.9 | 55.1 |
| | **South Asian** | 20 | 82.0 | 31.5 |
| **T1Q2** | **Caucasian** | 20 | 86.3 | 77.1 |
| | **South Asian** | 19 | 76.1 | 17.7 |
| **T1Q3** | **Caucasian** | 20 | 81.2 | 56.0 |
| | **South Asian** | 20 | 78.5 | 20.0 |

Descriptive Statistics for Response in Test 2

| | T2_examiner | N | Mean | SD |
|---|---|---|---|---|
| **T2Q1** | **Caucasian** | 20 | 74.7 | 17.0 |
| | **South Asian** | 20 | 73.7 | 17.9 |
| **T2Q2** | **Caucasian** | 19 | 75.4 | 15.9 |
| | **South Asian** | 20 | 67.5 | 15.6 |
| **T2Q3** | **Caucasian** | 20 | 74.5 | 20.2 |
| | **South Asian** | 20 | 73.7 | 20.5 |