THE ROLE OF NONVERBAL BEHAVIOR AND AFFECT ON RATINGS OF SECOND LANGUAGE PROFICIENCY

By

John Dylan Burton

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies — Doctor of Philosophy

2023

ABSTRACT

A long-standing problem in applied linguistics is how to account for nonverbal behavior in models of second language (L2) communicative ability (Canale, 1983; Canale & Swain, 1980; Celce-Murcia, 2007; Galaczi & Taylor, 2018; Hymes, 1972). Attempts have been made to incorporate some nonverbal behavior into some of these models, but they generally only account for strategic and interactional competences rather than the full range of information these behaviors can convey. It is well-established, however, that nonverbal behavior is fundamental to spoken, face-to-face communication (Hall et al., 2019; Hall & Knapp, 2013; Matsumoto et al., 2016), conveying semantic, cognitive, affective, and social-interactional information. Affect is one of the most important signals of nonverbal behavior, especially in facial movements (Knapp et al., 2013), conveying a range of emotive, orientational, and stance-related information. Nonetheless, language tests rarely account for this vital visual realm of information in their constructs or rating scales, despite research showing that it is meaningful to raters when formulating impressions of language proficiency (Choi, 2022; Ducasse & Brown, 2009; Jenkins & Parra, 2003; May, 2009, 2011; Nakatsuhara et al., 2021a; Nambiar & Goon, 1993; Neu, 1990; Orr, 2002; Sato & McNamara, 2019; Thompson, 2016). To date, few studies have observed measurable effects of nonverbal behavior or affect on impressions of language proficiency (Chong & Aryadoust, 2022; Kim et al., 2023; Nagle, 2022; Trofimovich et al., 2021; Tsunemoto et al., 2022).

To address this research gap, I designed a research study to triangulate ratings of affect, measurements of nonverbal behavior, and cognitive interviews to determine the role of nonverbal behavior and its affective information when listeners formulate impressions of language proficiency when observing L2 speakers' performances in a language test. I recruited 100 naïve, untrained raters to listen and watch short video recordings of 30 test takers interacting with an examiner during an oral proficiency interview in a high stakes speaking test. While listening and watching, the raters scored each speech sample on four categories of language—fluency, vocabulary, grammar, and comprehensibility—and ten categories of affect, covering dimensions of assuredness, involvement, and positivity. Following the rating activity, 20 of the raters took part in stimulated verbal recall sessions, which captured the raters' thought processes

while formulating their evaluations of L2 proficiency. In addition, I used automated, machine-learning software, iMotions, to extract measurements of nonverbal behavior in the speaking test samples in the form of engagement, attention, and valence. I also manually extracted speech and nonverbal behavior using multimodal annotations in ELAN.

The study broadly found that nonverbal behaviors and affect can impact proficiency outcomes in different ways. Desirable, communicatively-oriented behaviors such as mutual gaze, nodding, leaning forward posture, and representational gestures can convey confidence, engagement, and positive affect, which lead to differential outcomes in fluency, vocabulary, grammar, and comprehensibility. Comprehensibility, for example, was most impacted by raters' impressions of test taker engagement and also through behaviors that conveyed approachability, such as smiling and nodding. Fluency, vocabulary, and grammar were most impacted by impressions of confidence and low anxiety, as well as more target-like attentional focus. The raters were especially attuned to detecting listening comprehension, and the negative impact of comprehension breakdowns could be moderated when test takers took an adaptable stance that showed a desire to communicate. Overall, nonverbal behavior was found to affect perceptions of language proficiency in complex, dynamic ways that were mediated by interactions with the social context, requiring holistic interpretations of their impact.

By using naïve, untrained raters, this study has offered a glimpse into how non-linguists perceive language in real world settings. It thus has implications for language testing practice, as L2 speaking tests generally ask raters largely to ignore what they see and award scores based on what they hear. Speaking tests need to account for the nonverbal and affective repertoires of test takers in their constructs, rating scales, and rater training in order to capture test takers' full range of language ability. The raters' focus on listening comprehension also has implications for a broader adoption of integrated speaking assessments, where listening and speaking are assessed together. Finally, the study has important implications for how applied linguists conceptualize L2 communicative competence, as nonverbal behavior plays a much more important role than is currently ascribed.

Two languages cancel each other out, suggests Barthes, beckoning a third. Sometimes our words are few and far between, or simply ghosted. In which case the hand, although limited by the borders of skin and cartilage, can be that third language that animates where the tongue falters.
—Ocean Vuong

ACKNOWLEDGEMENTS

This dissertation wraps up a brilliant four year period at Michigan State University. I can safely say that I am walking away from this program with a new skillset and appreciation for knowledge that I did not have before. I have been continuously humbled by my mentors and teachers in this program that I have learned so much from. Even though my academic progress was rocked by COVID-19 in the second semester of my first year, the process of adapting and moving forward helped me pivot to this very project, which I may not have done otherwise.

First, I would like to recognize the various agencies that have so kindly funded or aided in this dissertation project. Without financial assistance, this project would have been much more taxing to carry forward. I would like to thank the International Language Testing System (IELTS) and Mina Patel for so kindly agreeing to provide the speech samples used in this study. These samples were an invaluable resource, and the inferences from this study are much stronger because of them. I would also like to thank the British Council Assessment Research Group, Duolingo, and The International Research Foundation (TIRF) for providing grants that allowed me to pay for participants, two research assistants, travel, and software for this project.

I cannot say how thankful I am to my advisor Professor Paula Winke. It really goes beyond words. From the day I started at Michigan State in May of 2019, Paula set out to socialize me into the world of academia. She taught me to become a better writer, a better research, a more critical thinker, and a steward for the field through editorial practices. Paula is absolutely a role model; she has shown unwavering positivity and dedication in her role as a mentor. I have learned about academic writing, grant and funding applications, departmental budgets, applying for jobs, and so much more. Perhaps most important, though, is how to be a mentor for others. I will take Paula's lessons with me as I move forward, and I hope I can provide her level of mentorship to others in the future.

I would also like to express my gratitude to my committee members. I feel extremely lucky to have been able to assemble such a stellar combination of experts: Dr. India Plough (language testing and nonverbal behavior), Dr. Aline Godfroid (SLA and psycholinguistics), Dr. Koen Van Gorp (language

testing and task-based language teaching and assessment), and Dr. Ryan Bowles (psychological assessment and statistics). The advice my committee provided at my dissertation proposal defense was indispensable, and this project is the result of those modifications. I owe a huge debt of gratitude to Dr. India Plough. As a result of interviewing India as part of a class project in LLT-861, I was motivated to study nonverbal behavior in the realm of language assessment. She kindly lent me her time and wisdom in an independent study on the topic during the worst part of the summer lockdown in 2020. I admire your curiosity and passion, and expertise in this topic area.

I would also like to thank so many others in the SLS program and at MSU. Dr. Shawn Loewen, Dr. Charlene Polio, and Dr. Peter De Costa have been outstanding teachers and colleagues to work with during these years. I am incredibly thankful for the kindness of Professor Monique Turner and her team at the CASE lab in the Department of Communications for allowing me to use iMotions in their lab. I would also like to thank SLS students and alumni for their ongoing comradery throughout the program. In particular, I would like to thank Robert Randez and Curtis Green-Eneix for allowing me to pilot my dissertation instruments with their students. An immense gratitude also goes to my research assistants Elena Gorshkova and Bethany Zulick, who were MA TESOL students when they assisted with my project.

Finally, I would like to thank a few others from my life outside of the program. I am grateful to the team at Lancaster University, notably my MA supervisor Professor Luke Harding and Professor Tineke Brunfaut, for their kindness and dedication, and for setting me on this path towards a PhD back in 2016. I am thankful for my colleagues at the British Council who have supported me over the years and taught me that work can also be about fun: Dr. Victoria Clark, Professor Barry O'Sullivan, Dr. Jamie Dunlea, Sheryl Cooke, and Mina Patel. I would like to thank my close friends Jeffery Walker, Amber Davis, Christopher Schaechtel, Matthew Brizzi, Jeremy Dickerson, Ryan Moltz, Katherine Macnair, and Matteo Cavazos for being supportive throughout this dissertation and listening to me talk about this project endlessly. Last but not least, thank you, Hollis Griffin, for your love, support, and understanding throughout the most difficult months of this writing up and completing this project.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

All life on earth dedicates sensory resources to perceiving the surrounding world. Sense enables life to preserve its existence by finding food, detecting threats, and reproducing. Life, however, does not exist in a solitary world. All beings, be they animals or even plants and fungi, interact with their world and each other by receiving and transmitting information in varying forms of communication. Communication may exist as words, sounds, signs, chemicals, smells, and possibly even electrical signals. Sense allows life to detect and decode information critical for survival. For humans and many animals, the primary modes of communication are visual and auditory, and meaning is conveyed and interpreted by hearing voices, sounds, barks, squeaks, and shrieks, and by seeing the direction of gaze, changes in posture, mouth movements, and other bodily actions.

Humans are unique in their ability to communicate very specific meaning by using highly complex, structured language. Language, either in the form of sounds or signs, communicates meaning that is generally symbolic, intentional, propositional, and purpose-driven (Buck, 1984). The use of language and choice of words can be automatized, but ultimately language choices are made by speakers with a certain degree of awareness. The body, however, is the existential foundation of human culture and sense perception (Bourdieu, 1977). Individuals pick up on stances, feelings, and orientations by seeing the body language of others. This body language, or nonverbal behavior, may convey a wide range of information about traits and states, and may also align with language to strengthen certain meanings or emphasize information. Communication occurring through nonverbal behavior is thought to largely occur spontaneously, automatically, and without attention, and, unlike linguistic forms, is also considered to be largely non-propositional and unbound from form-meaning relationships (Buck & VanLear, 2002). That is to say, a person's hand movement making a rising motion may indicate a rising action, but it could also indicate any manner of other relationships to other words or ideas. Interactant listeners receive these multiple channels of information and decode a wealth of information from speakers, ranging from desires and needs to intuition about underlying feelings and goals. Nonverbal cues may enhance comprehension of these various lines of information in these interactants, thus serving as the "co-text" for the communicative

event (Rost, 2016, p. 42).

One defining feature of nonverbal behavior is its usefulness in communicating information about speakers' personalities and affective *stances* (Burgoon et al., 2016; Knapp et al., 2014). These affective stances may include emotions, feelings, moods, attitudes, and orientations towards others, all of which are critical for communicators when establishing and maintaining interpersonal relationships. These affective stances furthermore combine with pragmatic acts to maintain (or disrupt) social harmony (Brown & Levinson, 1987; Roever, 2021) Nonverbal behavior allows us to monitor conversations and the impact of our words on others. We can often see when someone is happy by observing their smiles. We can also see if they are upset if we see them looking away and frowning. Interpreting the affective stances of others allows people to predict what someone might say or do, and it can clear up ambiguities when the verbal message might be interpreted in multiple ways (e.g., a smirk when the verbal message is intentionally ironic). The communication of affect may also be "contagious," and emotions may be spread through nonverbal channels (Elfenbein, 2014; Hatfield et al., 1994). One only needs to think of a room of children in which one child begins laughing and suddenly the room erupts in laughter. These nonverbal channels may also synchronize between interactants (Hess & Fischer, 2013), as evidenced by people adapting similar gestural patterns together, or the co-occurrence of smiles in conversation. Where the broad interpretation of language is through semantics, the interpretation of nonverbal behavior is largely affective-interactional.

If communication is largely holistic, comprising interwoven threads of verbal and nonverbal communication (Burgoon et al., 2016), it is unknown whether speaking skills can be assessed independently of what a person sees during speech. This is an important question for individuals involved in the assessment of speech, such as those that teach presentation skills, speech pathologists, and individuals that work with second language development. Workers in these fields are interested in the abilities of people to speak in their first or second language, and assessors may need to produce scores that provide meaningful inferences about these abilities. It could then be possible, for example, that affective stances observed through nonverbal behavior may subtly alter interpretations of language ability. Alternatively, it may be the case that the most accurate interpretations of language ability may only be possible by considering both verbal

and nonverbal channels of communication together. Understanding how the visual world interacts with speech is then critical for the field of speech assessment.

Second language assessment

The communicative turn in applied linguistics (Firth & Wagner, 1997; Halliday, 1985) involved a paradigmatic shift towards assessing productive skills, that is, speaking and writing, as previously these skills were rarely assessed or assessed indirectly (Fulcher, 2003). Nowadays, using performance assessments has become the standard in most language testing programs. These assessments take a range of formats, but they all share a common defining characteristic:

> [A]ctual performances of relevant tasks are required of candidates, rather than more abstract demonstration of knowledge, often by means of pencil-and-paper tests… The *format* of a performance-based assessment is distinguished from the traditional assessment by the presence of two factors: a *performance* by the candidate which is observed and judged using an agreed *judging process*" (McNamara, 1996, pp. 7–10, emphases original).

This judging process has generally involved the use of human raters who award scores based on a set of predetermined and empirically validated criteria, which are often used operationally in the form of a rating rubric. Despite decades of research and notable improvements in the reliability in the use of human raters, the scoring of performance assessments is still an active area of research (Knoch et al., 2021) due to its centrality in arguments about fairness in language testing (Kunnan, 2018; McNamara et al., 2019).

Within the domain of performance assessment rating, a great number of researchers have conducted studies that investigated characteristics that may have an impact on test scores, such as test-task characteristics (e.g., the question type or question difficulty level), test taker characteristics (e.g., gender, first language, age), and rater characteristics (e.g., nationality, professional/educational background). The rationale for these studies is often to uncover potential sources of bias. "Bias occurs … when large numbers of items systematically and demonstrably (dis)advantage specific populations on construct-irrelevant grounds, such as gender, educational background, or home language" (Deygers, 2019, p.10). Not all impact sources, however, qualify as score bias. When construct-relevant aspects of tests affect scores—such as

grammatical accuracy and complexity—the discussion instead surrounds the relative impact of particular performance features. These factors *should* influence scores. In assessment contexts, it is critical to document variance to ensure that scores meaningfully align with the intended construct. If variance is found due to factors unaccounted for by rating scales, this evidence can support the revision of scales or rater training programs (or both) where needed, as it can be evidence of construct underrepresentation (Messick, 1989). Desirable variance can lead to elements being added to scales, and undesirable variance (construct irrelevant-variance, Messick, 1989) can lead to elements being removed or changes in rater training programs to address those problematic features.

For many types of speaking tests, behavior elicited through tasks is multimodal; that is, spoken language is accompanied by the test takers' visible nonverbal behavior. A rater may conduct a test with one or more test takers, listen to their language, and produce a score about their second language ability. Nonetheless, most rating criteria only include descriptions of performances that are purely language-based, even when the purported test construct is based on communicative ability rather than linguistic accuracy or complexity. Raters may then be expected to narrow their focus and ignore any salient visible information—often implicitly—and award scores only from speech. Research has suggested, however, that raters notice the nonverbal behavior and affective stances of test takers (Ducasse & Brown, 2009; May, 2009, 2011; Sato & McNamara, 2019), which may ultimately impact scores beyond differences in language ability. Humans may not be able to ignore salient aspects of communication to focus only on the verbal mode of communication. This could also partially explain why raters can be idiosyncratic, formulating their own internal rating processes (e.g., Lumley, 2002). Despite repeated calls for research in this area (Pennycook, 1985; Kellerman, 1992; Plough et al., 2018; Plough, 2021; Young, 2002), there is still limited information about which types of performances are most impacted by visual information, the direction and size of this impact, and whether nonverbal behavior may impact all rating criteria equally. Understanding the nature of this impact will shed light on how nonverbal behavior fits into models of language proficiency that generally only tacitly acknowledge nonverbal communication.

Aims of this dissertation

This dissertation contributes to the ongoing discussion of nonverbal behavior in the second language literature. By studying nonverbal communication in the context of language assessment, it is possible to consider speech perception from the rater's perspective, identifying elements of the visual realm that raters take into account when forming an impression of language proficiency. This dissertation is thus a study of the following key topics:

1. The relationship between measures of nonverbal behavior and the more linguistic constructs of fluency, vocabulary, grammar, and comprehensibility
2. The relationship between judgements of affect—one channel of information from nonverbal behavior—and the more linguistic constructs of fluency, vocabulary, grammar, and comprehensibility
3. The salience of particular nonverbal behaviors during the rating process
4. The interpretation of nonverbal behavior when arriving at judgements of language proficiency

The dissertation is divided into eight chapters including the introduction.. Chapter 2 presents background literature on this topic, specifically detailing popular models of second language (L2) communication and how they incorporate extra-linguistic elements, the roles and relationships of nonverbal behavior with language, and the origins and impact of interpersonal affect. Chapter 3 presents the research questions drawn from the literature review, followed by Chapter 4, which covers the methods and organization of this research project. Chapters 5, 6, and 7 will present separate analyses based on analytical methods used, the first two quantitative, and the last a qualitative analysis. Chapter 8 will discuss and synthesize findings, and Chapter 9 will conclude with the overall contributions of this study, limitations, and paths forward.

CHAPTER 2: LITERATURE

Researchers of L2 speech assessment have sought to identify measurable features of spoken language that provide inferences about the overall communicative language ability of learners. Crucial, then, is an understanding of the structure of communicative language ability to select features that can be reliably measured, yet give a full, nuanced picture of what someone can do with language. Once decisions are made about the dimensions speech will be measured upon, validation research must then provide evidence of the integrity of those ratings and their meaningfulness. Studies of this type, however, have repeatedly shown that raters attend to far more than just the linguistic features of speech. They attend to a wide range of phenomena, including the nonverbal behavior produced during the test, as well as the affective stances interpreted from these behaviors. The literature to date has painted a picture of behavior and affect influencing speaking test scores, but there is still a dearth of research available that indicates the size and extent of this impact, and which constituent features matter most to raters.

This literature review will be organized into three key sections. First, I review literature related to the construct of L2 proficiency. This will include a discussion of models that proficiency is broadly based on, such as communicative competence, communicative language ability, and interactional competence. I will illustrate how nonverbal behavior and affect have been acknowledged but systematically underrepresented in these models of communication. Underrepresentation will be discussed in studies in which raters described features that define communicative success. I will provide evidence that modality matters in the rating of performances, with audiovisual speaking tests consistently resulting in higher scores than audio-only tests, showing that variance in tests scores is unaccounted for when raters can see the test taker. The second section will review nonverbal behavior. I will discuss the semantic, cognitive, social, and affective information that behavior can convey, how nonverbal behavior relates to language, and how it is tied to culture. The section will end with a discussion of how nonverbal behavior impacts interpersonal perceptions, with a detailed review of the literature pertaining to language testing. I provide evidence that nonverbal behavior forms an important part of the fabric of what raters attend to. Finally, in the third section, I will discuss in greater detail the affective function of nonverbal behavior. I discuss definitions of affect,

emotions, and feelings, and provide evidence that the affective function of nonverbal behavior is simultaneously a cognitive, social, and cultural phenomenon. I discuss its relationship to language achievement in studies largely in SLA research, and finally I provide evidence that affect can have important implications in interpersonal relationships when it is interpreted by listener-raters. As a core function of nonverbal behavior, affect can also impact language proficiency scores. The review will conclude with a short section on the measurement of nonverbal behavior and affect.

The construct of second language assessment

Second language proficiency

Speaking is an everyday skill. It is used to communicate quickly, solve problems, and engage with others. When people learn languages, the first skill that comes to mind is often speaking. The cultural norm in English is not to ask *Do you read any other languages?* but rather *Do you speak any other languages?* Speaking is thus at the heart of how society views L2 competence. When it comes to speech assessment, however, it may be the most difficult skill to assess (Fan & Yan, 2020) given its ephemeral, complex, socially contextualized, and dynamic nature. For this reason, speaking was often neglected in favor of testing more discretized aspects of language such as reading comprehension (Fulcher, 2003).

At the heart of any discussion of language assessment is a discussion of the *construct*. Tests aim to measure aspects of language, and the construct provides a theoretical orientation to what is being measured. Most contemporary speaking tests claim to measure second language proficiency, which is a construct that draws on frameworks of communicative competence (Celce-Murcia, 2007; Hymes, 1972; Canale & Swain, 1980; Canale, 1983), communicative language ability (Bachman & Palmer, 1996, 2010), and, sometimes, incorporates elements of interactional competence (Galaczi & Taylor, 2018; Kramsch, 1986; Plough et al., 2018). These frameworks lay the groundwork for what it means to communicate effectively. Local, regional, and national language development standards likewise draw from these same frameworks in their categories and descriptions of language development. Some examples of these are the Common European Framework of Reference (CEFR; Council of Europe, 2020), World-Class Instructional Design and Assessment (WIDA; WIDA, 2020), Interagency Language Roundtable (ILR; Interagency Language Roundtable, n.d.), and the

American Council on the Teaching of Foreign Languages (ACTFL; The National Standards Collaborative Board, 2015). Each test or set of standards covers varying aspects and amounts of the underlying construct of L2 speaking ability, but due to its complexity, it is impossible to cover all constituent aspects. In fact, researchers in SLA argue that many of these models are missing cognitive components of speech prossessing and production (e.g., Levelt, 1989; Levelt et al., 1999) and prediction (Levinson, 2016; Pickering & Garrod, 2013) that characterize various stages of acquisition (Hulstijn, 2015), as psycholinguistic aspects of speech processing are critical to language proficiency (de Jong, 2023).

De Jong (2023) presented perhaps the most comprehensive description of language proficiency to date, containing psycholinguistic, structural, and socio-interactional elements. Psycholinguistic elements of speech processing (Levelt, 1989, 1999) include the following cognitive subskills:

> a skill to conceptualize the preverbal message, a skill to retrieve the correct lexical items quickly along with their morphosyntactic characteristics, a skill to retrieve the appropriate sounds with these lexical items and to plan them as connected speech, a skill to send motor programs to the articulatory muscles to produce intelligible sounds, and finally, skills to efficiently monitor one's speech. (de Jong, 2023, p. 542)

De Jong also stressed a central role for interlocutor input comprehension and prediction as key cognitive elements (Levinson, 2016; Pickering & Garrod, 2013), as these elements are critical predictors of quick, spontaneous, fluent speech. The automatization of the processes mediating comprehension, prediction, and production is what characterizes L2 fluency (Kormos, 2006). Structural elements of speech include elements of linguistic competence drawn from Bachman and Palmer (1996, 2010), Canale and Swain (1980), and Canale (1983). These include grammatical, lexical, and phonological competencies. These hierarchical models also included discursive, pragmatic, and strategic components of speech, but for reasons I explain below, I reimagine these as social-interactional elements. Finally, de Jong's (2023) final grouping of elements of language proficiency are those that have an outwardly social-interactional focus. These include competencies that allow learners to mediate between their core linguistic knowledge and the outside world, requiring a knowledge of contextual appropriateness and audience, and requiring the ability

to negotiate and co-construct meaning among conversational interactants. These elements stress interactional competence (Kramsch, 1986; Galaczi & Taylor, 2018; Young, 2011), or "the ability to listen attentively, to design the message for the recipient…, to manage the conversation, and to use appropriate nonverbal behavior" (De Jong, 2023, p. 544).

De Jong (2023, p. 545) presented a hierarchical model of these elements, drawing from and adding to models from Bachman and Palmer (1996, 2010). In her model, language proficiency is made up of linguistic competence and strategic competence. Linguistic competences include structural (grammar, vocabulary, phonological forms), predictive, and pragmatic (sociolinguistic and functional) competences. Strategic competences, the other main branch in her model, include self- and other-supporting mechanisms and planning. However, the model is lacking in its explanatory power as there is little distinction between the cognitive or social orientation of competences. Purely cognitive aspects of speech (prediction), for example, are listed alongside structural and organizational features, and social-interactional forms are subsumed within discourse competence. If prediction, comprehension, and production form elements at the core of language use, a reformulated, nested model may better represent language proficiency. Drawing from de Jong's (2023) descriptions, Figure 2.1 may better represent a sociocognitive model of language proficiency.

Figure 2.1
A Sociocognitive Model of Language Proficiency (Adapted From de Jong, 2022)

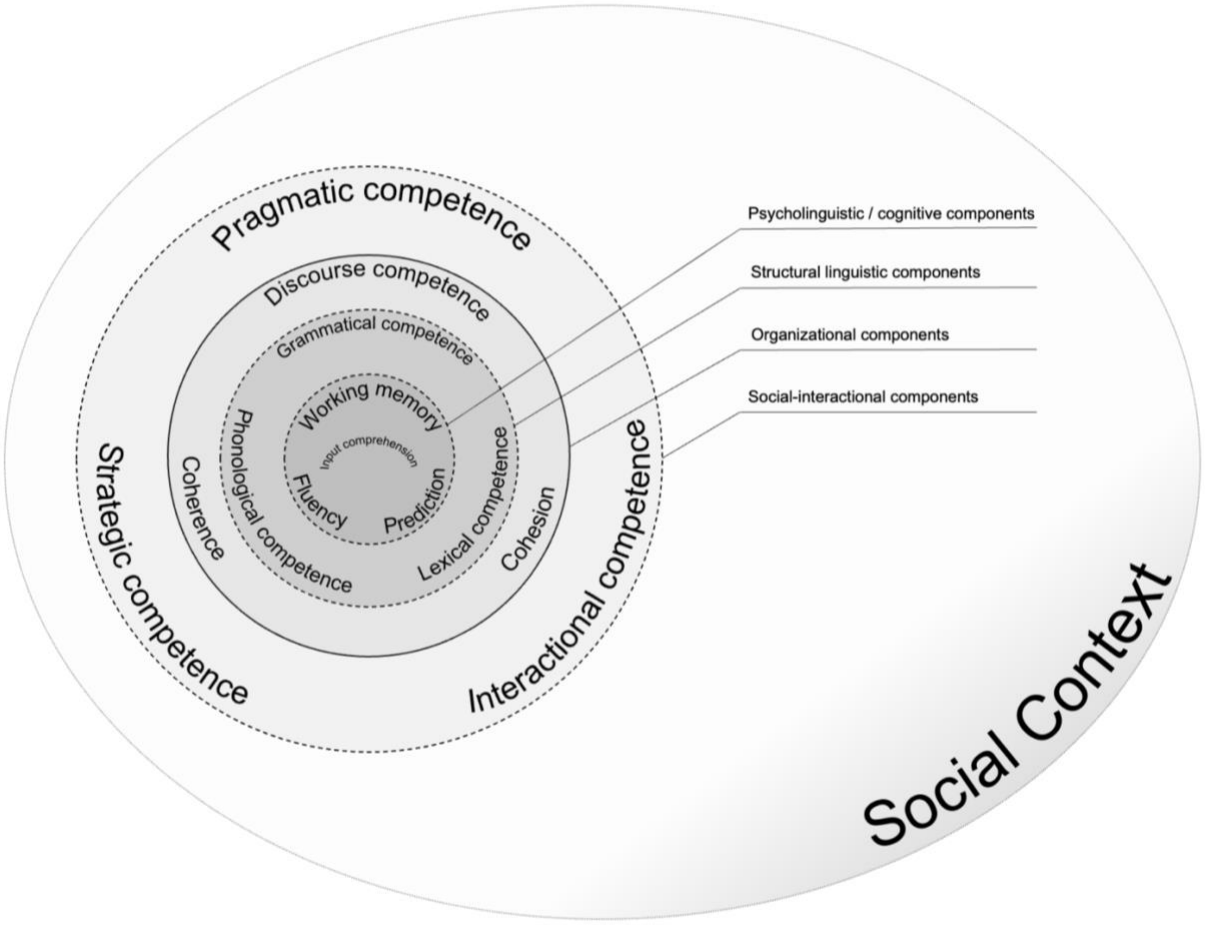


Figure 2.1 is organized in such a way to stress a gradation between individual, psycholinguistic components at the center, and socially oriented elements at the edge. The boundaries between psycholinguistic elements are not strict, as indicated by dashed lines. Components within competences are not listed as discrete units, as they represent constructs that overlap to some degree (e.g., grammar and lexis also exist as lexicogrammar; pragmatic speech acts are often interactional in nature). Psycholinguistic and cognitive elements of language use are at the core of language proficiency. These elements exist primarily within the individual speaker and interact with structural components to produce speech (Levelt, 1993, 1999). As speech is proceduralized and automatized through use and practice of the language, comprehension and prediction become faster, enabling cognitive fluency to develop (Segalowitz, 2010).

Thus, the core aspects of production are strengthened. Structural elements, which also exist within the individual, include grammar, lexis, and phonology, and form the bulk of linguistic competence (Bachman & Palmer, 1996, 2010; Canale & Swain, 1980). These begin as declarative knowledge, and as speakers access them repeatedly they are proceduralized and eventually become part of their automatic language use (DeKeyser, 1997, 2020). Thus, there is a close interrelationship between structural and cognitive competences. Discourse components, at a higher level than lexicosyntactic structural components (see e.g. Celce-Murcia, 2007), allow speakers to connect structural forms to assemble meaning-making units of language. De Jong (2023) argued that discursive and interactional competences are closely linked, as they draw on similar linguistic forms. All discursive and interactional elements draw on lexical and syntactic forms, but the way speech is organized at a micro level, independent of context (which I define as discourse competence here) helps the speaker convey a coherent and cohesive message. The knowledge and ability to produce coherent sentential and intratext structure in speech and writing is an element that I envisage as primarily psycholinguistic and within the speaker, while the ability to manage conversational exchanges draws more heavily on context and is socially dependent, and thus an aspect of interactional competence. For this reason, I have used a bolded line around discourse competence, as these components are largely within the cognitive realm of speakers' abilities. Structural, psycholinguistic, and discourse components of speech represent the majority of what is assessed in many contemporary language tests, such as the rating categories fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation in the International English Language Testing System (IELTS; IELTS, n.d.). These three core units represent what is traditionally thought of in psycholinguistic constructs of language proficiency.

Social-interactional components, on the outside of the sphere, orient towards meaning-making with others. They allow speakers to use their core language comprehension, knowledge of structures, organizational skills, and production skills to navigate communicative contexts. These social-interactional components are other-focused, and aid in the co-construction of meaning (Young, 2011), and for this reason, their external boundary is also dashed, as it indicates that meaning is bound with the social context. Social-interactional components—including pragmatic, strategic, and interactional competences—go beyond

structural forms and fluency. For example, within pragmatic competence, the sociolinguistic context requires changes in structural forms in order for the user to choose, for example, the right level of register or politeness (e.g., Brown & Levinson, 1987; Roever, 2021). Pragmatic competence furthermore allows speakers to choose the correct functional language in a social situation (e.g., greeting someone instead of saying goodbye). Any deviation in appropriate register or function can result in unintentional breakdowns in social meaning. Similarly, strategic competence enables speakers to manage the creation of their messages by planning utterances, moderating their own speech, and self-repair when necessary. Strategies are meaning-focused and exist to support communication. Interactional competence, as described above, then helps speakers organize conversations with others. The use of these social-interactional components is mediated by the needs of the greater social context, surrounding it. These three groups (cognitive, structural, and social-interactional), interact with each other in speech, but may be better seen as organized by these cognitive-social dimensions. Nonetheless, these boundaries exist mainly as a visual metaphor, as in reality all aspects of speech may be context-dependent.

Important to this discussion, however, is the extent to which nonverbal behavior and/or affect are included in definitions of language proficiency, and if they are, how they are conceptualized. De Jong's (2023) description only explicitly mentions nonverbal behavior in the context of interactional competence. However, other frameworks have included nonverbal behavior in their models of communicative competence as well. The following section will review how applied linguists have conceptualized the role of nonverbal behavior and affect within their models. Following the literature and the dissertation study, I will revisit whether de Jong's model in Figure 2.1 adequately represents L2 communication.

Behavior in models of communication

There are many terms from the literature describing L2 communicative ability. Communicative competence (Canale, 1983; Canale & Swain, 1980; Celce-Murcia, 2007; Celce-Murcia et al., 1995; Hymes, 1972), communicative language ability (Bachman & Palmer, 1996, 2010), and language proficiency (Hujlstein, 2015) have all been used to describe broadly similar frameworks that detail components of successful communication. Models of interactional competence (Galaczi & Taylor, 2018; Young, 2011),

rather than full communicative models, describe essential interactional skills to communication that exist alongside communicative competence. This list is not exhaustive either, as many others have written on the topic, though these are the most influential. Other frameworks, such as (interpersonal) communication competence (Morreale et al., 2013) and intercultural communicative competence (Byram, 2021) extended the notion of communication to other sociocultural contexts. The Common European Framework of Reference (CEFR; Council of Europe, 2020) is a framework of L2 communicative language development across multiple scales and subscales related to real world language use. Most of these frameworks acknowledge the presence of nonverbal behavior, each with different functional attributes ascribed to it. In these descriptions, rather than recounting in full the models of each author, I will instead focus on how they operationalize nonverbal and affective behavior as aspects of communication.

Communicative competence. Hymes (1972) developed a theory of communicative competence at a time when there was heavy debate over differences between competence and performance (cf. Chomsky, 1965). His development of communicative competence incorporated grammatical competence and contextual/sociolinguistic competence, and he further argued that competence was dependent on both knowledge and the ability to use that knowledge in context. While linguistic knowledge was seen as result of learning various components of a language (grammar, vocabulary, phonology, and organization features), "ability for use" was defined as the functional abilities of individuals to deploy their linguistic knowledge. He described ability for use as comprising cognitive, volitive, and affective factors, with affective components (such as motivation) partially determining an individual's level of competence. Importantly, he also makes the case for the inclusion of an interactional domain of competence, drawing from the work of Goffman. Hymes (1972) also explicitly included nonverbal behavior as part of his model:

> In both respects the interrelation of knowledge of distinct codes (verbal: non-verbal) is crucial. In some cases these interrelations will bespeak an additional level of competence… Within the view of communicative competence taken here, the influence can be expected to be reciprocal. (p. 284)

This quote suggests that Hymes viewed nonverbal communication as contributing meaning across various aspects of language use and not comprising any separate functions on its own. Canale and Swain (1980),

however, took a more restrictive view of the role of nonverbal communication.

Canale and Swain (1980) set out to model Hymes' (1972) ideas with the goal of drafting, essentially, a formalized construct of communicative competence for teaching and testing. They stressed the inclusion of sociolinguistic aspects of competence along with grammatical (syntactic, lexical, and phonological) and strategic (compensatory) competences, as these are critical to real and authentic language use. However, they maintained a separation with communicative performance, which included "the relationship of these competences and their interaction in the actual production and comprehension of utterances (under general psychological constraints that are unique to performance)" (p. 6). Canale and Swain (1980) thus explicitly tease out the cognitive and affective aspects of Hymes' (1972) competence. Nonverbal behavior was granted a role only in strategic competence, which was the "verbal and nonverbal communication strategies that may be called into action to compensate for breakdowns in communication due to performance variables or insufficient competence" (p. 30). They described these as largely compensatory (e.g., paraphrasing), though they also included some underdefined aspects of sociolinguistic competence when interacting with others, where these are seen as coping mechanisms. Nonetheless, they acknowledged in 1980 that nonverbal behavior may one day be given a larger role in models of L2 ability:

> More research on the role of such nonverbal elements of communication as gestures and facial expressions in second language communication may reveal that these are important aspects of communication that should be accorded more prominence in the theory we have adopted. (p. 36)

Canale (1983) reformulated their model slightly to include discourse competence (the organization of language into coherent textual units), again assigning nonverbal behavior a compensatory role, though also acknowledging its ability to enhance communicative effectiveness such as "deliberately slow and soft speech for rhetorical effect" (p. 10).

Celce-Murcia (2007), drawing from her own past work (1995; Celce-Murcia et al., 1995), attempted to extend Canale and Swain's (1980) and Canale's (1983) models with a focus towards language pedagogy. She proposed a model with four key subcompetences. This model included a reformulation of some of the labels of Canale and Swain's (1980) subcomponents (linguistic and socio-cultural competence),

but also added in a new component called *actional competence*, described as the ability to understand and produce pragmatic speech acts and sets. She also included interactional competence alongside the other three competences and included strategic and discourse competence as uniting competences mediating the other four. Important to this discussion, however, is that she is perhaps the first author to provide a more nuanced and specific vision of nonverbal behavior in a larger conceptualization of communicative competence. She described a non-verbal/paralinguistic competence embedded within interactional competence, with bodily behaviors (gestures, gaze behaviors, nodding, and other body language), proxemics (orientation in space), haptic behavior (touching), and paralinguistic cues (non-verbal sounds) as core components. However, these behaviors were only allocated roles within interactional competence.

Other authors also speculated about the inclusion of nonverbal behavior in discussions of communicative competence. Scarcella et al. (1990) added a component of "verbal, non-verbal, and paralinguistic knowledge underlying the ability to organize spoken and written texts meaningfully and appropriately" (p. 72). Almost identical to Celce-Murcia (2007), however, these were given almost exclusively interactional turn-taking roles. Likewise, Savignon (1972) and Jakobovits (1970), both contemporaries of Hymes in the early 70s, also speculated on the importance of nonverbal behavior to communicative competence. As will be seen later, however, there is evidence for nonverbal behavior at nearly all levels of communication.

Communicative language ability. Bachman and Palmer (1996, 2010) extended the work of Canale (1983) and Canale and Swain (1980) to develop a conceptual framework of L2 language ability, which they termed communicative language ability. Language ability for them was derived from complex interactions between the sociolinguistic context and language use. Language use, "the creation or interpretation of intended meanings in discourse by an individual, or as the dynamic and interactive negotiation of intended meanings between two or more individuals in a particular situation" (p. 61), is essentially language performance as described by Canale and Swain (1980). Language use was described as being made up of language knowledge and strategic competence, which would then interact with topical knowledge, affective schemata, and characteristics of the language use situation to produce meaning.

Affective schemata here are the "affective or emotional correlates of topical knowledge" (Bachman & Palmer, 1996, p. 65), and are characteristics of a task that may provoke an affective response in the speaker due to past emotional experiences. These affective responses can then facilitate or hinder the test taker's ability or willingness to communicate in a given situation depending on the response's emotional weight. The ability to discuss affect is included in their list of pragmatic competences, called "knowledge of ideational functions" (p. 70). The ability in question uses linguistic resources to explicitly *discuss* affect (e.g., "I'm angry", "I'm disappointed in the results"), rather than leveraging multimodal resources to *express* affect (e.g., nodding, directing attention to speaker, and verbally backchanneling to show engagement). Their presentation of strategic competence, which is a broadened view of metacognitive strategies rather than mainly compensatory strategies, does not include an explicit discussion of nonverbal behavior. Overall, Bachman and Palmer (1996) saw the danger of agitating test takers by using emotionally charged content, but they did not outright discuss any impact of nonverbal behavior on the communication of affect or interaction, as they viewed speaking as a primarily audio-based skill.

Fourteen years later, Bachman and Palmer (2010) largely reaffirmed their previous conceptual framework, while giving a greater role to the interactional effects of their affective schemata. Affective schemata were still seen as mostly task-based emotional weights, with an additional caveat added about individuals' orientations to interpersonal communication. However, in their descriptions of affective schemata in interaction, the authors explicitly invoked the use of nonverbal behavior when encoding and decoding meaning. For example, when describing a customer service encounter, a waiter:

> engages his affective schemata when [the customer] reacts to how he seems to be feeling about the conversation—he looks relaxed and not impatient about taking her order… She also takes into account her affective schemata—does she feel confident enough about her ability to speak Thai to participate in a conversation or is she so nervous that she only feels up to pointing to items on the menu. (p. 39)

Implicit here is that both the customer and waiter are encoding and decoding each other's nonverbal signals to arrive at a conclusion about the communicative intent of the other individual. Nonetheless, the authors

only ascribed verbal characteristics of speech to the language abilities used to convey meaning. Similar to their 1996 framework, Bachman and Palmer (2010) did not place much emphasis on nonverbal behavior in their formulation of strategic competence, though they did mention that it can play a role in the appraisal of communication:

> In a conversational exchange, the language user can appraise the extent to which his communicative goal has been accomplished by the way his interlocutor responds, with language, with non-verbal communication, or with both… Affective schemata are involved in determining the extent to which failure was due to inadequate effort, to the difficulty of the task, or to random sources of interference. (p. 53)

Nonverbal behavior then appears to have a minor role in navigating coping mechanisms, though not as clearly as with the compensatory strategies in Canale and Swain (1980).

Interactional competence. As a reaction to knowledge-based, accuracy-focused language proficiency assessments, Kramsch (1986) proposed *interactional competence* as a socially oriented test construct that taps into speakers' abilities to manage interaction and create meaning in conversational contexts. Interactional competence is "the ability to co-construct interaction in a purposeful and meaningful way, taking into account sociocultural and pragmatic dimensions of the speech situation and event" (Galaczi & Taylor, 2018, p. 226). It consists of abilities that allow speakers to systematically manage turn-taking, repair, topic management, and agreements with sensitivity toward the listener and the interactional context (Hellermann, 2008; Pekarek Doehler & Berger, 2018). Dating back at least to the time of Hymes (1972), interactional abilities were seen as relying heavily on both verbal and nonverbal channels. This was later reaffirmed in Celce-Murica's (2007) framework, where a range of different behaviors were specified in their relationship to interaction. With Canale and Swain (1980), nonverbal behavior related to strategic competence in regard to compensatory strategies and coping mechanisms, which may be seen as similar in nature to repair strategies in conversation. Frameworks of interactional competence have been developed over the years to detail the range of features speakers employ to manage conversational interactions, and in these, authors have consistently provided a place for nonverbal behavior. Nonetheless, it has been argued

that the specification and implementation of nonverbal behavior in interactional assessments still needs further work (Plough et al., 2018; Plough, 2021; Roever & Kasper, 2018).

Galaczi and Taylor (2018) provided the most visual representation of interactional competence to date, bringing together macro- and micro-level features underlying the construct from a broad reading of the literature. Their interactional competence "tree" (p. 227) visualizes spoken interaction as derived primarily from social context, which defines the speech event and the speech act. These social-contextual variables determine how speakers and listeners interact with each other and how they deploy their interactional abilities. As an example, in a formal social context with a hierarchical power dynamic, one can imagine fewer initiations and turns granted to listeners with lower power status, while in a more informal, balanced social context, there will be more balance in how conversations are co-constructed. Arising from social context then are the various functions of interactional competence, seen as branches, while micro-level forms such as initiating and closing a conversation are listed as leaves. The functional categories of interactional competence included turn management, topic management, breakdown repair, and interactive listening, along with their constituent micro-level features. They also listed a specific place for non-verbal behavior as a function, including features of facial expressions, laughter, and posture.

This conceptualization is somewhat problematic, however, since nonverbal behavior is not characterized by propositional, form-function relationships (Buck & VanLear, 2002), as instead it co-occurs with speech to accomplish the various functional categories of interactional competence (e.g., achieving intersubjectivity, Burch & Kley, 2020; repair, Burton, 2021a; maintaining progressivity, Hırçın Çoban & Sert, 2020).

Other models. There are other models of L2 communication that also deserve mention in this review, as these include to some degree mentions of the roles of affect or nonverbal behavior. Van Ek (1986) developed a model of communicative ability with an emphasis on both communication skills and also personal and social development. In his model, he added to Canale's (1983) components of linguistic, sociolinguistic, discourse, and strategic competence two additional categories: sociocultural and social competence. Sociocultural context here referred to the familiarity with cultural norms in the communicative

context and the ability to navigate these. Social competence included elements of interactional competence, but also aspects of affect, namely motivation, attitude, self-confidence, empathy, and the ability to handle social situations (van Ek, 1986, p. 65). Nonetheless, despite the inclusion of an ability to handle affect when navigating social situations, nonverbal behavior was not mentioned explicitly.

Byram (2021) drew on van Ek's (1986) work in his development of an extension of communicative competence to address cross-cultural contact in language teaching. His framework of intercultural communicative competence included van Ek's (1986) components, but stressed the various skills, knowledge, education, and attitudes necessary for individuals from different cultural backgrounds to engage in an "effective exchange of information" and "establishing and maintaining relationships" (p. 43). His attitudinal domain of intercultural communicative competence made a case for the reduction of bias in order to respect "people who are different in respect to the cultural meanings, beliefs, values, and behaviors they exhibit" (p. 44). Essential here is navigating affect, both one's own and the perceived affect from others. This category also included an implicit focus on nonverbal behavior, though he was wary of any explicit inclusion in this model due to prescriptivism of native speaker norms:

> [Nonverbal behavior] is clearly an element of interaction which is crucial, but the challenge to the dominance of the native speaker as model applies just as much here as it does to standards of verbal communication. In other words, any teaching of non-verbal skills and knowledge should enhance competences as an intercultural speaker, not imitation of a native speaker" (Byram, 2021, p. 59).

Morreale et al. (2013) discussed a framework of "communication competence" removed from a focus on language teaching or L2 developmental models. Their notion of competence, essential to communication in any language, was defined as "the extent to which people achieve desired outcomes through behavior acceptable to a situation" (Morreale et al., 2013, p. 25) and was dynamically defined by context. This model hinged on the perceptions of others, as communicative success is interactional in nature: "how we actually behave in most instances is less important than how others *perceive* us to have behaved " (Morreale et al., 2013, p. 25, emphasis in original). Their model consisted of three main elements—motivation, knowledge, and skills—nested within context. Motivation included the communicative

objectives and goals for a communicative act, as well as affective components driving or inhibiting these acts. Knowledge consisted of the content or procedural information relevant to successful communication, including topics, semantics, and discursive functions of language. Skills included the behaviors central to communication itself, including macro-level (functional) and micro-level (form) skills. Skills explicitly included verbal and nonverbal macro- and micro-level skills. Finally, culture provided the framing of communicative events, including cultural, relational, and situational types, as well as interpersonal, public, and mediational levels. Morreale et al.'s (2013) framework drew from the work of Burgoon (2016, first published in 2010) to be the broadest description of the functional role of nonverbal behavior within a larger model of communicative competence: "People use nonverbal behavior to complement verbal messages, to regulate interactions, and to define the socio-emotional quality of relationships" (p. 104). They delineated the specific roles of gesture, eye gaze, posture, facial movements, paralinguistics, haptics, proxemics, and chronemics, although largely describing their affective output. This model, however, is meant for general communicative success (e.g., in an L1) and not to describe L2 use or development, but it provides a useful contrast to models developed in applied linguistics.

The Common European Framework of Reference (CEFR) (Council of Europe, 2020) also provides an example of how nonverbal behavior is operationalized in a set of standards concerning language ability. The CEFR includes illustrative scales that detail language development across multiple domains of communication. It is organized principally by skills (reading, writing, speaking, listening), interaction, mediation, and strategies. It also includes scales of linguistic, sociolinguistic, pragmatic, plurilingual, and pluricultural competence. As such, it is to date the most complete view of a developmental understanding of L2 communication. The CEFR does include nonverbal behavior as well, but it is generally only limited to descriptions at very low levels (Pre-A1–A1) in a limited number of subscales. For example, in the category Overall Mediation, the A1 descriptor is:

> Can use simple words/signs and non-verbal signals to show interest in an idea. Can convey simple, predictable information of immediate interest given in short, simple signs and notices, posters and programs. (Council of Europe, 2020, p. 92).

Descriptors at the A2 level and above do not include a continuation of the development in nonverbal behavior. Similar descriptors can be found in subscales of Leading Group Work, Facilitating Pluricultural Space, Mediating Concepts, and Mediating Communication. One subscale that deviates from this paradigm is Strategies to Explain a New Concept. There are no descriptors for A1–A2 levels, with the following starting at B1:

> Can make a set of instructions easier to understand by repeating them slowly, a few words/signs at a time, employing verbal and non-verbal emphasis to facilitate understanding. (Council of Europe, 2020, p. 120).

No other descriptors above the B1 level further delineate a cline of nonverbal skills within this subscale. Another scale that deviates from the prior paradigm is that of Interaction within Qualitative Features of Spoken Language scales. Here, nonverbal behavior is described only at the highest level, C2:

> Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. (Council of Europe, 2020, p. 183).

Given that nonverbal cues are generally encoded and decoded automatically (Gifford, 2013), it is unclear which behaviors this descriptor refers to. Similar to the other scales, there are no descriptors below the C2 level illustrating a development in nonverbal behavior within this category.

Interestingly, the inclusion of nonverbal behavior within the CEFR roughly follows patterns from Hymes (1972), Celce-Murcia (2007), and Galaczi and Taylor (2018) in its inclusion within categories of interaction, or Canale and Swain (1980) and Canale (1983) in its inclusion in a category of strategies. Nonetheless, these inclusions are generally at a very low level, and descriptors offering a view of development within nonverbal patterns at different ability levels are not given for any of the subscales apart from signed languages, which are beyond the scope of this discussion.

The CEFR does, however, include substantially more information about the encoding and decoding of affect. For example, the Conversation scales describe how a learner "can express how they feel in simple terms" at A1, "can express and respond to feelings such as surprise, happiness, sadness, interest, and indifference" at B1, "can convey degrees of emotion" at B2, and "can use language flexibly for social

purposes, including emotional, allusive, and joking usage" at C1 (Council of Europe, 2020, pp. 73–74). Other categories that cover affect to varying degrees are Reading Correspondence, Correspondence, Online Conversation and Discussion, Expressing a Personal Response to Creative Texts, and Sociolinguistic Appropriateness.

Rater reports of behavior in studies of communication

The development of models of L2 communication is generally top-down. That is to say, these models are generally theoretical, developed by scholars based on readings from the literature and their own observations or intuition about language. On the other hand, a bottom-up approach can be taken by asking raters to listen to L2 speech and describe the various features that comprise communicative effectiveness. The features they notice and are able to describe can then be used as validation evidence for models, and any features they notice that are not included in models may be then included in further revisions. Understanding what untrained, linguistic laypeople describe is important when designing tests of L2 ability because:

> the ultimate arbiters of L2 speakers' oral performance are typically not in fact trained language professionals, who have meta-level linguistic insight and are possibly concerned primarily with features of communication that are the focus of their own training as linguists or language teachers, but interlocutors with no specialist training. (Sato & McNamara, 2019, p. 895)

Thus, the views of these individuals can help refine models and test constructs.

Early research on oral proficiency assessments consistently showed that raters attend to a range of linguistic and nonlinguistic criteria when scoring, such as features of content and discourse management (Halleck, 1992; Lazaraton, 1996; Neu, 1990; Ross, 1992; Young & He, 1998). There are few studies, however, that have considered a broad range of performance features that raters notice, particularly through rater reports, when orienting towards L2 communication ability. Orr (2002), for example, used trained raters when eliciting speech features that factored into successful performances on the First Certificate in English speaking test (a test linked to the B2 level on the CEFR). The raters commented on the linguistic criteria present in the scales they were trained on, but also noted a range of 12 other characteristics, such as content-

related task features, exertion/effort, test preparation, and nonverbal behavior. The authors did not explain how raters oriented towards nonverbal behavior in the study. Brown et al. (2005) also investigated the criteria raters used when rating the TOEFL iBT. Similar to Orr (2002), raters attended primarily to the linguistic criteria given in the rating scales, but the content of speech relating to idea development and task success was the next most important category raters mentioned. Nonverbal behavior was not attended to given that the TOEFL iBT is rated audio only, with no visual material present.

Sato and McNamara (2019) also elicited untrained raters' internal scoring criteria in an attempt to reveal underlying factors impacting impressions of communicative effectiveness. They had 23 novice raters (postgraduate students without knowledge of applied linguistics) view and rate twenty speech samples on the speakers' overall communicative effectiveness. The speech samples were drawn from performances on the College English Test-Spoken English Test (CET-SET) delivered in China, and dyadic interactions from Cambridge Assessments. Raters used a seven-point holistic scale of communicative success with endpoints of poor and excellent. Afterwards, raters provided stimulated verbal recalls and retrospective interview data supporting their decisions. The raters discussed linguistic features of communicative success most often, followed by a sizeable number of comments regarding general communicative success (categorized as general comments relating to task success, comprehensibility, etc.). Content-related features made up the third greatest number of comments, echoing Orr (2002) and Brown et al. (2005). Orientations to nonverbal behavior and affect made up the next most sizeable number of features. These were categorized separately but discussed together. The final category of comments related to interaction. While the specific findings related to nonverbal behavior and affect will be discussed in detail later, important here is to note that raters oriented to aspects of communication outside the traditional realm of communicative competence, as "their judgements of communicative ability are based on a wider range of speech features and speaker behaviors than the constructs of current proficiency tests" (Sato & McNamara, 2019, p. 911). In this sense, communicative success as measured by the CET-SET and Cambridge Assessment tests could suffer from construct underrepresentation, as content-, nonverbal behavior-, and affect-related features are not present in these testing organizations' rating scales and could thus "potentially misrepresent the judgements of real-

world interlocutors" (p. 912).

Similarly, Ducasse and Brown (2009) and May (2011) asked raters to report salient features of successful and unsuccessful interactions. Ducasse and Brown (2009) asked 12 listeners with teaching backgrounds to watch 17 pairs of beginning-level Spanish students taking a discussion-based test. Raters were asked to watch and then record retrospective impressions of the paired interactions, and following this, they watched the video a second time and provided stimulated verbal recalls on what constituted interactional abilities. Interaction, notably, was not defined for the raters. The raters primarily oriented to non-verbal interpersonal communication, followed by interactive listening (comprehension and support), interactional management (topic management and coherence). In a similar design, May (2011) asked four trained raters to discuss the performance features of 12 intermediate and advanced level learners taking paired speaking tests. Raters rated the tests in pairs using an analytic rating scale focusing on both linguistic and interactional aspects of language. Afterwards, raters provided retrospective reports, discussion recordings with the other rater scoring the sample, and interviews with the researcher commenting on features that were salient when scoring interactional effectiveness. She found that raters focused on whether the dyads understood each other's messages, responded to each other, and used communication strategies. Both nonverbal and affective aspects of interaction surfaced in the responding category, along with listening comprehension, comprehensibility, and other aspects of interactional competence (e.g., turn and topic management, repair):

> The ability to work together cooperatively, manage a conversation, communicate with assertiveness, demonstrate effective body language and interactive listening, and thus help to co-construct a collaborative pattern of interaction were regarded by the raters as key aspects of a successful interaction. (May, 2011, p. 140)

Important to note is that raters focused on these criteria despite the fact that they were not present in the original rating scale, again supporting Sato and McNamara's (2019) claim that raters tapped into a larger set of skills and abilities than delineated by the test construct.

Modality effects

Ratings in different delivery modes, namely audio-only and audiovisual rating, have the potential to uncover whether the visual world as a whole may exert an impact on ratings. If differences do exist, then *something* in the visual world, be it the presence of nonverbal behavior, its interpretation via affect, or some other explanation must be the driving force of those differences. To date, research in this area has rather convincingly found that scores based on audiovisual tests—that is, where the rater can see the test taker—are higher than scores where raters only hear the speech of the test taker (audio-only) (Choi, 2022; Conlan et al., 1994; Gullberg, 1998; Larson, 1984; Nambiar & Goon, 1993; Nakatsuhara et al., 2021; Styles, 1993). These effects are group-level effects, however, as individual test takers may have equivalent scores in both modes or even occasionally lower scores in the audiovisual format. Only one study found lower scores in the audiovisual mode (Lavolette, 2013), while others found roughly equivalent scores (Beltrán, 2016; Thompson, 2016; Shohamy, 1994; Uludag et al., 2022). Because these studies that found conflicting results were for the most part underpowered, they will not be discussed here. I will discuss the three most definitive studies on this topic in turn: Choi (2022), Nakatsuhara et al. (2021a), and Nambiar and Goon (1993).

Nambiar and Goon (1993) were one of the first to speculate on the possible impact of modality differences in language tests. The authors had raters conduct speaking tests with 87 undergraduate students and score them face-to-face, after which the same samples were rated by the same raters as audio-only recordings. The speaking test include two tasks, of which one was in an interview format with the rater, and the second was a paired discussion task. The study found that mean scores based on audio recordings were significantly lower than the ones rated in the face-to-face mode, with the mean audio-only score dropping 1.25 points in the interview task (out of 20 points) and 0.47 points in the dyadic task (also out of 20 points). Students in the first quartile of scorers (the highest scorers) were impacted by the audio-only mode difference more negatively than students in the bottom quartile (the lowest scorers). The researchers noted that raters found interpreting pausing, silence, and grammatical/phonological inaccuracies in the audio-only format to be difficult to interpret without visual information. Raters were better able to understand the source of breakdowns in fluency in the face-to-face rating and were less attuned to inaccuracies. However,

the authors noted the limitation that the raters also served as interlocutors in the speaking test design, which may have played a role in their scoring tendencies.

Nakatsuhara et al. (2021a) investigated the effect of mode on International English Language Testing System (IELTS) scores in three scenarios: live, audio-recorded, and video-recorded. Using Many-Facet Rasch Measurement (MFRM) on 6 raters' scores, they found that audio-only rating resulted in scores that were overall 0.92 logits more difficult than the video rating mode, resulting in a half band difference in final scores after rounding (where band scores were ordinal units on a 9-unit rating scale). An analysis of the individual criteria showed the same trend, with all four criteria (fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation) marked lower in the audio-only mode, and lexical resource was impacted the most. Nonetheless, these trends were not consistent across all raters, as one rater in particular was found to be biased negatively towards video rather than audio. An analysis of the raters' verbal protocols revealed that in many of the cases, the raters scored the audiovisual samples higher because they helped examiners "a) to understand what the test takers were saying, b) to comprehend better what test takers were communicating using non-verbal means …, and c) to understand with greater confidence the source of test takers hesitation, pauses, and awkwardness" (Nakatsuhara et al., 2021, p. 19).

Finally, Choi (2022) investigated the score differences of 110 test takers on two asynchronous audiovisual recordings conducted on Zoom (with and without the interlocutor present) and asynchronous audio-only recordings. Eight trained raters scored these samples in an anchored dataset, and the results were analyzed with confirmatory factor analysis and MFRM. She found that the data did not support a one-factor model, but instead a three-factor model where the three delivery modes each represented separate latent variables with variance attributable to different sources. In other words, visual elements in the video-recorded samples represented differing relationships between the test scores and the higher-level latent variable of L2 proficiency, which may have expanded the construct of speaking to include nonverbal behavior. Regarding score differences, supporting previous studies, she found that audio-only scores were lower than both video-recorded formats, and the two video formats were approximately equivalent. Audio-only ratings were approximately 0.5 logits more difficult than the video recording with an interlocutor

present, and 0.75 logits more difficult than the video recordings with the interlocutor removed. Similar to Nakatsuhara et al. (2021a), these differences represented approximately a half-band difference on the same 9-point rating scale.

Overall, these studies point to key differences between audio-only and audiovisual based ratings. Audio-only scores were found to be consistently lower than audiovisual scores, with differences representing about half a band score in Nakatsuhara et al. (2021a) and Choi (2022), which used the same rating scale. In these studies, raters appeared to judge linguistic criteria in largely consistent ways between the two rating designs, but the presence of visual information helped them to make a more informed assessment of the test taker's language ability, resulting in these score differences. However, as Nakatsuhara et al. (2021a) and Choi (2022) noted, not all raters behaved in the same way. Some were influenced more by visual information than others, and some outright ignored its presence and rated each sample equivalently. Interpreting language proficiency in light of visual information may thus be idiosyncratic, despite the presence of group-level effects. Nonetheless, all authors hypothesized that score differences due to the presence of video were likely the result of the impact of nonverbal behavior.

Summary

The studies reviewed so far all suggest that nonverbal behavior and affective responses during test scenarios play a much larger role than ascribed by models of language proficiency (de Jong, 2023), communicative competence (Canale, 1983; Canale & Swain, 1980), and communicative language ability (Bachman & Palmer, 1996, 2010). Nonverbal and affective behavior play critical roles in test discourse (Ducasse & Brown, 2009; May, 2011; Sato & McNamara, 2019), and they are important meaning-making devices that possibly contribute to variance in speaking test scores (Choi, 2022; Nakatsuhara et al., 2021a; Nambiar & Goon, 1993). The evidence to date points to the fact that these behaviors indeed form a critical aspect of Hymes' (1972) ability for use. Yet, despite wide consensus that nonverbal behavior is an essential aspect of speaking, it has been neglected in models of L2 communication and in speaking test constructs (Plough, 2021). In the next two sections, I consider nonverbal behavior and affect separately. In the section on nonverbal behavior, I review the various functions of nonverbal behavior and how these have been

shown to impact speaking test scores. An important question in this section is whether the role of nonverbal behavior is limited to interaction or whether it may also play a role in perceived linguistic competence, which indeed appears to be the case from the studies by Nambiar and Goon (1993), Nakatsuhara et al. (2021a), and Choi (2022). The section on affect will be arranged similarly, detailing the meaning of affect and its impact during interaction, as well as how it may relate to language learning outcomes and ratings on language tests.

Nonverbal behavior

Nonverbal behavior is fundamental to spoken, face-to-face communication (Hall et al., 2019; Hall & Knapp, 2013; Matsumoto et al., 2016). It is always present in spoken communication, predates language, develops prior to verbal ability in infants, and precedes verbalizations in interactional encounters (Burgoon et al., 2016). Nonverbal communication combines with verbal communication in the encoding and decoding of meaning in speakers and listeners (Halberstadt et al., 2013). Complex multimodal Gestalts (Mondada, 2014)—patterns of linguistic constructions and nonverbal behavior in sociocultural contexts—can also inform observers of underlying cognitive, psychological, or (socio)affective states of speakers (Argyle, 1988; Guerrero & Wiedmaier, 2013; Schmid Mast & Cousin, 2013). Nonverbal behavior also plays an important role in conveying the content of speech in conversation (Kendon, 2004; McNeill, 1992, 2005). Nonverbal behavior has been studied in thousands of articles in fields as disperse as neuroscience, psychology, sociology, anthropology, computer science, engineering, robotics, medicine, communication, and applied linguistics (Plusquellec & Denault, 2018). However, despite the critical mass growing that recognizes the central role of nonverbal behavior to communication, it has been somewhat neglected in linguistics and applied linguistics—and perhaps especially language testing—as these fields stress the autonomy of language within communication. Mondada (2016), drawing on Derrida's (1967) critique of the Western overreliance on speech as a source of truth, called this restricted focus *logocentric*, in contrast to an embodied view of language:

> Producing talk involves visible breathing and articulating movements not only of the face and the mouth, but of the entire body; moreover, these articulatory movements are dissociable from other

bodily conduct… both talk and gesture originate from the same process. (p. 340).

To account for the full range of communicative skills, she argued that applied linguists should adopt a less logocentric view of language to incorporate a wider range of embodied meaning. Regarding language development, Stam (2008) remarked that "looking at learners' gestures and speech can give us a clearer picture of their proficiency in their L2 than looking at speech alone" (p. 253). If the visual realm is informative about language proficiency, it is thus also important to study the visual realm in the context of language testing and assessment.

Setting boundaries on what constitutes nonverbal behavior, however, is a complex endeavor. Although Hall et al. (2019) defined nonverbal communication as "a behavior of the face, body, or voice minus the linguistic content, in other words, everything but the words" (p. 272), others have included other aspects relating to the social context. Argyle (1988), for example, identified eight dimensions of nonverbal communication, including facial expressions, gaze behavior, gestures, posture, haptics (touch), proxemics (spatial location and movement), appearance (clothing and other objects), and paralinguistics (nonverbal sounds such as laughing or sighing). Chronemics (Walther & Tidwell, 1995), or the ways in which individuals use or perceive time, can also impact how messages are interpreted, such as when the same message is conveyed at different times of the day. Vocalics—the expressiveness and animation of the human voice which may or may not be grouped with paralinguistics—can also impact how people are perceived in terms of affect and likeability (Mehrabian, 1981). For the purposes of this paper, I will restrict the focus to that of Birdwhistle's (1970) kinesics: facial expressions, gaze behavior, gestures, posture, and other bodily movements. I will refer to these broadly as nonverbal behavior.

The literature on nonverbal behavior is vast, too vast to cover in any one dissertation. Indeed, entire volumes, handbooks, edited volumes, textbooks, and journals have been produced to tackle the various research areas on the topic. For this reason, I will not break down this literature review by nonverbal form (e.g., smiling, gaze shifts, etc., but see Givens & White, 2021 for an exhaustive treatment on the subject), but rather by the role behaviors play and how these roles relate to L2 use and development. The categorization of nonverbal behavior has a long history but was first seminally codified in Ekman and

Friesen's (1969) five categories: emblems (codified signs), illustrators (abstract meanings), affect displays, regulators (interactional signs), and adaptors (non-meaning making movements). This section will cover similar categories but in terms of function, namely the semantic, cognitive, social-interactional, and socio-affective information conveyed by nonverbal behavior, as well as its cultural origins. I will then turn to how nonverbal behavior factors into interpersonal perceptions in language testing and assessment. The focus here will be on rater-test taker interactions, as these can be a source of score variance.

Semantic information

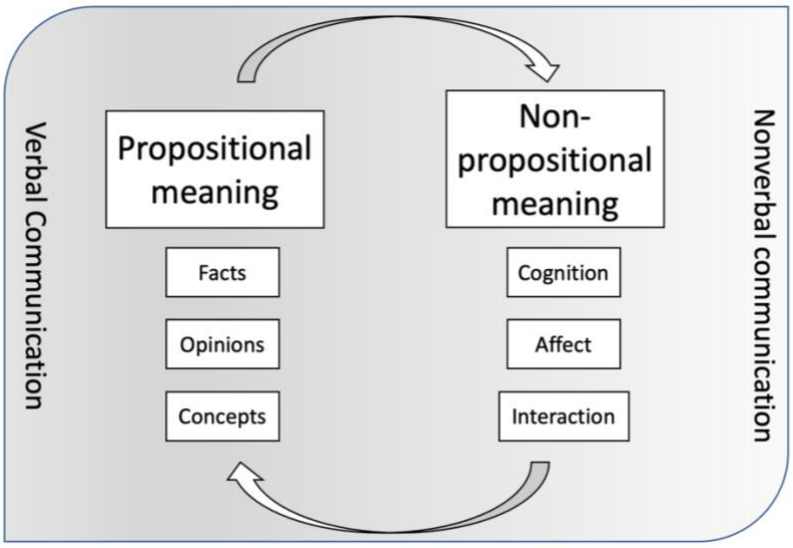
Dual-process models of language hold that two streams of information make up communication: one is acquired, intentional, propositional, and symbolic; the other is ontogenic, automatic, non-intentional, non-propositional, and made up of signs (Buck, 1984). Communication can then convey discrete, constructed meanings and implicit, almost unnoticeable information that occur simultaneously and interact in all social contexts, developing over the course of one's life (Buck & van Lear, 2002). Generally, verbal language is propositional in that statements can be logically negated. Utterances can be negated ("the plane took off") by using evidence available to the listener (the plane is on the ground). Verbal language is symbolic as it is made up of units (spoken or signed) that generally exhibit a bound form-meaning relationship with discrete meanings. The ability to produce these meanings is learned. That is to say, although words may be understood in many ways, the broad possibilities of meaning are restricted to a defined number of possible interpretations at one particular moment in time (e.g., "cool" may indicate a lower temperature, or an informal statement of approval, but it is unlikely to mean "feline"). Nonverbal behavior, on the other hand, is generally of the spontaneous category. Information encoded by speakers in the face, eyes, sound of the voice, or body may be unintentional, but may reflect ongoing internal or social processes. Listeners may decode this information implicitly and unknowingly. These behaviors are often non-propositional in that they cannot be negated; the surprise that someone appears to show as a plane takes off carries no logical statement to negate (though the perceived emotion may be questioned; "don't act surprised!"). They are non-symbolic in that a gestural flourish may not have any specific meaning the

speaker intended to convey. The flourish may then be a metaphor for the verbal content of the speaker.

Although verbal information is conveyed in an almost entirely symbolic, intentional way, nonverbal information can be either symbolic/intentional or spontaneous/unintentional (Kendon, 2004; McNeill, 1992, 2005). For example, if a teacher asks a student a question ("Did you see the plane take off outside?") and the student either nods their head or gives a thumbs up emblematic gesture, this information is conveyed intentionally and symbolically. The nod or thumbs up gesture both convey agreement; they are intentional, and they can be negated if the teacher in fact knows that the student did not see the plane take off. Likewise, if the student is describing the plane taking off and simultaneously uses a hand movement in a rising fashion, this representational gesture adds intentional and symbolic meaning to the verbal utterance. If the student said "the plane landed" while using the same rising gesture, there would be a breakdown in the propositional meaning of the combined information. The student may also use other symbolic, intentional gestures when speaking to manage the interaction with the teacher as well. However, alongside these symbolic moves will be unintentional, non-propositional information conveyed by the student's face, hands, and body. The student may show an expressive, excited reaction in the face and use an upright, engaged posture. These behaviors will color the informational, event-added affective information to their verbal utterances. Gestures, in particular, "may reveal systems or richer underlying distinctions than are apparent in speech alone… that is, semantic distinctions not apparent in speech may instead appear in gestures" (Gullberg, 2022, p. 321). In other words, the information nonverbal behavior conveys is on a cline rather than a strict dichotomy between propositional and non-propositional. The two feed into the same system (Buck & van Lear, 2002), and influence the interpretation of the overall message. A visualization of this relationship is presented in Figure 2.2, which I designed drawing from my understanding of the literature presented above. The ombre color in the background suggests that both verbal and nonverbal

information can convey both types of meaning, and the arrows suggest that they feed into each other.

Figure 2.2
The Relationship Between Propositional and Non-propositional Meaning



The example of the "thumbs up" gesture showed that nonverbal behavior can convey meaning that is propositional, intentional, and symbolic; one category of such behavior is *emblematic* gestures. These gestures are those that have a codified symbol and interpretation in a given sociocultural context (Kita, 2009; Morris et al., 1979). Examples include the hand balled into a fist with the thumb raised (thumbs up), the hand closed with the index and middle finger raised (peace sign), and the palm facing up with all fingers touching together (in Italian, *mano a borsa*). The first two are common in American English, while the second, if turned so that the palm faces the speaker, is pejorative in British English. The third is commonly used in Italian for emphatic purposes. These gestures have meanings that are culturally bound and shared by particular groups of individuals. Some studies have shown that learners orient towards these culturally bound L2 gestures, facilitating the acquisition of language (Allen, 1995) and also the gestures themselves (Belío-Apaolaza & Hernández Muñoz, 2021).

However, most gestures are not as rigid in their form-function relationship, belonging to the category of spontaneous communication. These are not codified or lexicalized, yet they sometimes still provide semantic information through lexical or syntactic forms. Some spontaneous gestures fill linguistic functions, such as identifying referential content in deictic expressions by filling syntactic structural slots

(saying, for example, "Go!" while pointing at the bedroom) or substituting for particular pragmatic speech acts (waving and smiling, to greet someone) (Gullberg et al., 2010). Others, such as the hand raising while describing a plane taking off or motioning up with the head, reinforce the lexical-semantic content by providing a visual representation through motion events (Cadierno, 2004; Choi & Bowerman, 1991). These *iconic* gestures (gestures closely aligned with semantic content) always co-occur with speech (McNeill & Duncan, 2000; Graziano & Gullberg, 2018) and synchronize at phonological, semantic, and pragmatic levels (McNeill, 1992). Spontaneous, co-speech gestures have the same meaning as the speech utterance, and they have the same pragmatic functions (Kendon, 1980; McNeill, 1985, 1992). McNeill (1992) and others hypothesized spontaneous gesture and speech as a single integrated system (Clark, 1996; Engle, 1998; Goldin-Meadow & Alibali, 2013; Kendon, 2004), where nonverbal behavior provides imagery in a "language-gesture dialectic" (McNeill, 2005, p. 25).

An example of the lexicosyntactic encoding of gesture is with motion events. Languages encode syntactic and lexical information in how they indicate manner and path motion differently (Slobin, 2006; Talmy, 1985, 2000), and languages may vary in how gestures align with path motion in speech (e.g., Özyürek & Kita, 1999; Slobin, 1996). For example, verb-framed languages encode directionality within the verb itself (e.g., Spanish *subir* (go up), while satellite-framed languages such as English encode directionality on a unit such as a particle (e.g., *up* in go up; *go* has no directional meaning). Spanish speakers then coordinate gestures corresponding to path on verbs (Negueruela et al., 2004; Stam, 2006), while English speakers tend to encode directionality on the satellite (McNeill & Duncan, 2000; Slobin, 1996; Stam, 2006). Studies analyzing acquisition have found that there is some evidence of transfer in the position of target-like path gestures in learners of typologically different languages from their own (Gullberg, 2009a, 2009b; Stam, 1998, 2017) as well as other iconic gestures (McCafferty & Ahmed, 2000), though even highly proficient speakers may remain entrenched in their L1 gestural patterns (Choi & Lantolf, 2008; Stam, 2008, 2010). Thus, it is possible that some types of nonverbal behavior conveying semantic information can be acquired. The bulk of research on semantic aspects of nonverbal behavior has involved gesture because of its special relationship to speech, but less is known about the semantic aspects and/or acquisition

of other forms, such as nodding and head shaking.

In Cienki's (2012) integrated view of language and behavior, spoken language is the primary mode for the generation and communication of ideas; all other modes—including but not limited to gesture, facial expressions, paralinguistics—take on symbolic and communicative roles depending on contextual needs. The boundaries of language are then flexible, with context determining whether a mode contributes additional meaning. For example, in a telephone call, the body conveys no meaning whatsoever, while in an emergency, face-to-face setting, the face and body may convey the majority of information necessary to interpret the severity of the encounter. Meaning, in Cienki's (2012) view, can be migratory depending on needs. One can agree with an interactant by saying so, by using an emblem (thumbs up), or by nodding their head. In teleconferencing, for example, hands are generally less visible due to the limited viewing angle of the camera. If an interactant desires to show agreement in a group of several others, they may avoid speaking as this would interrupt the call and purposefully show their hands on the screen in a thumbs up gesture or nod emphatically, which may be less common in face-to-face settings (Mark et al., 2023). In this framework, "a family of meanings is thus dynamically paired with a family of forms" (Morgenstern & Goldin-Meadow, 2022, p. 6). Cope and Kalantzis (2020) and Kalantzis and Cope (2020) developed a framework for a multimodal grammar that takes into account the transitional, migratory nature of meaning across modalities, including speech, body, sound, space, and even images and text. For them, meaning is unbound from modality and may transition from one to another depending on the constraints defined by context.

Cognitive information

Various theories exist regarding the link between gesture, speech, and thought. The majority of these posit that all three are linked, but they differ in the degree to which speech or gesture is an integrated (McNeill, 1992, 2005) or co-orchestrated system (Kendon, 2004, 2007; Goldin-Meadow & Brentari, 2017). As mentioned earlier, McNeill (1992, 2005) and McNeill and Duncan (2020) theorized that some nonverbal behaviors—in particular, spontaneous gestures—originate in the same cognitive processes as speech. In their growth point theory, gestures (and perhaps other behaviors) thus offer glimpses into cognition (Goldin-

Meadow & Alibali, 2013), and gestures may lighten the cognitive burden of the production of speech (Cassell et al., 1999; Goldin-Meadow et al., 2001). The interface hypothesis (Kita & Özyürek, 2003), which also considered gestures and speech as interlinked, considered the interplay between visual and linguistic thought. Other theories considered behavior as a facilitator of lexical retrieval (a compensatory system), such as in the lexical retrieval hypothesis (Krauss et al., 2000), or as instrumental in the construction and representation of visual thought to be verbalized in the information packaging hypothesis (Alibali et al., 2000). Despite the different ways each theory connects behavior and cognition, all theories account for some degree of integration:

> Co-speech gestures, which are often produced without conscious awareness, are synchronous with speech, cannot be understood independently of speech, perform similar pragmatic functions as speech, and are multifunctional in that they perform both cognitive and communicative functions often at the same time. (Stam & Tellier, 2022, p. 336)

Perhaps because of this very tightly integrated system, gestures even parallel breakdowns during speech disfluencies in L2 speakers (Graziano & Gullberg, 2018; Seyfeddinipur, 2006).

There may be different cognitive mechanisms that influence behavior, such as affect, context, and language proficiency. As will be discussed in coming sections, affective states can result in varying autonomic and behavioral responses in speakers, such as anxiety causing changes in averted gaze, relative expressiveness, and a higher number of self-adapting behaviors (Gregersen, 2005; Lindberg et al., 2021, 2022). Some cognitive mechanisms may be context-dependent, such as pupil dilation increasing as a reflection of greater task difficulty (van der Wel & van Steenbergen, 2018) and stimuli familiarity (Heaver & Hutton 2011, Otero et al. 2011). Underlying L2 ability, of interest to this study, may also impact differential production of gestures. In bilinguals, speakers may gesture more in their less dominant, weaker language (Aziz & Nicoladis, 2018; Benazzo & Morgenstern, 2014; Gullberg, 2006, 2012; Krauss & Hadar, 1999; Nicoladis, 2007; Nicoladis et al., 2007), though these results have been contested (Gullberg, 1998; Laurant & Nicoladis, 2015; Nicoladis et al.; 1999; Sherman & Nicoladis, 2004). In a study of 75 Spanish language learners at beginner, intermediate, and advanced levels, Gregersen et al. (2009) found that learners

differed from the previous findings in their gestural output depending on proficiency level. They found fewer illustrators—co-speech-occurring gestures that enhance the speaker's meaning—in video recordings of speakers with lower proficiency, while more advanced speakers gestured more often in meaning-enhancing ways: "By using more illustrator gestures, advanced learners reinforced grammaticality, used visual discourse markers, strategically reinforced meaning through the visual channel, and, in general, responded with sociolinguistic gestural dexterity" (Gregersen et al., 2009, p. 205). They also found that learners in beginner and intermediate levels used more self-adapting gestures, such as hand fidgeting and adjusting clothing, than more proficient speakers. There were no group differences in the use of compensatory gestures used to convey meaning when lexical retrieval was delayed or resulted in a breakdown. Lin (2022) found similar results in Chinese speakers of L2 English, with a greater number of illustrators and beats (gestures indicating phonological stress and rhythm) at advanced levels, while less proficient speakers used a greater number of deictic gestures (pointing) and compensatory movements. Learners with differing proficiency profiles may also differ in the abstractness or concreteness in which they use deictic gestures (i.e., pointing) when specifying semantic referents (So et al., 2013).

Behavior can also affect speech comprehension and production by enhancing the perception, interpretation, reactivity, and memory storage of utterances (Beattie & Shovelton, 1999; Cohen & Otterbein, 1992; Drijvers & Özyürek, 2017; Graham & Argyle, 1975; Holler et al., 2018; Kelly et al., 1999; Tellier, 2008), as well as in L2 speech (Hardison & Pennington, 2021; Morett, 2014). Neurocognitive studies have shown that the brain processes and decodes the two modes in similar ways (Özyürek & Kelly, 2007; Özyürek, 2014). Co-speech gestures may facilitate a reduction in the load on limited working memory resources when speaking (Cook et al., 2012; Krauss et al., 2000), and serve compensatory roles when verbal resources are limited (Frick-Horbury & Guttentag, 1998; Hosetter & Alibali, 2007). Gesture, when restricted, can also lead to the increased production of dysfluencies or otherwise limited language use in the weaker language or L2 (Graham & Heywood, 1975; Laurant & Nicoladis, 2015; Rauscher et al., 1996). When cognitive load is increased, such as when listening to a difficult question, speakers may orient away from the visual input available to them by averting their gaze (Doherty-Sneddon & Phelps, 2005; Doherty-

Sneddon et al., 2002; Glenberg et al., 1998); this can free up cognitive resources, then allowing the speaker to provide an answer to the question. For example, in Burton (2023), greater proportions of averted gaze were found to be the result of increased question difficulty in an online L2 speaking test. Although speakers looked away from the camera/interlocutor more when preparing their answer to more difficult questions, question difficulty had no relationship with blinking frequency. In these studies of gaze directional changes, the task or affective changes due to task difficulty influenced behavior, which was then hypothesized to benefit cognition.

Facial behavior can also have unconscious physiological effects on listeners. There is evidence that seeing spontaneous communication in the faces of others creates natural neurocognitive linkages between interactants (Morris et al., 1996; Suslow et al., 2006); this happens when brain states reach a type of unity between speakers that exerts a powerful influence on emotions and social organization (Buck & Powers, 2006). In some hypotheses, the brain interprets behavior in others and, to some extent, attempts to replicate it in the listener, such as the active intermodal matching hypothesis (Meltzoff & Moore, 1997) or the mirror neuron system (Rizzolatti & Craighero, 2004). Viewing certain facial configurations can lead to differing cardiac (Levenson & Ekman, 2002) or respiratory responses (Boiten, 1996), perhaps due to the tight connection between the face and emotion (Levenson et al., 1990). In short, behavior can have a reactive effect in listeners, resulting in unconscious behavioral changes (such as synchrony, mimicking, or possibly aversion) that may originate at a neurological level (Dimberg et al., 2000).

Nonverbal behavior can also provide important information to listeners during conversational exchanges. It can facilitate listening comprehension in L1-L1 encounters (Drijvers & Özyürek, 2017; Goldin-Meadow, 2003; McGurk & McDonald, 1976) as well as in the context of L2 speakers or listeners of L2 speech (Dahl & Ludvigsen, 2014; Nakatsukasa, 2016; Sueyoshi & Hardison, 2005; Tsunemoto et al., 2022). Visible gestures may help listeners predict or infer about the semantic nature of physical objects speakers are thinking about before these are verbalized (Pine et al., 2010). Listeners may furthermore interpret nonverbal behavior to infer cognitive, social, affective, or trait information about speakers, often completely automatically, unconsciously, and extremely quickly, within microseconds of social

interactions (Ambady, 2010; Borkenau et al., 2009; Lakin, 2006). When forming an impression of whether a speaker understood certain questions, listeners may orient towards speakers' gestures (Goldin-Meadow et al., 1992) or facial behaviors (McDonough et al., 2019, 2022b, 2023). More information about interpersonal encounters and how nonverbal communication contributes to state or trait judgements is presented in later sections.

Social-interactional information

Interaction is dynamic, spontaneous, and sequentially organized, and transitions in interactional turns between speakers are organized in observable ways (Sacks et al., 1974). Meaning is co-constructed by interactants through the exchange and reciprocation of ideas (Young, 2011), and in dialogue, speakers simultaneously convey meaning as well as other interactional information about their performance, understanding, and intentions (Clark, 2002). As discussed in the section on L2 communication and language proficiency, nonverbal behavior has been recognized for its role in the management of social interaction in studies of L2 communication and interactional competence (Celce-Murcia, 2007; Dai, 2023; Galaczi & Taylor, 2018; Hymes, 1972; Kramsch, 1986; Plough et al., 2018; Scarcella et al., 1990). The ability to manage social interactions—interactional competence—is essentially one of organization and pragmatic function, as it describes how speakers assign turns, how these turns and actions contribute to coherent meaning making, how breakdowns are repaired so that all speakers can follow the conversation, and how specific units of interaction such as openings and closings are organized (Schegloff, 2006). Interactional behaviors convey embodied semiotic information that facilitates communication, yet do not have direct lexicosyntactic functions (Gregersen et al., 2009; Gullberg, 1998). Along with verbal resources, these behaviors allow speakers to reach, restore, and maintain intersubjectivity (Burch & Kley, 2020; Goodwin, 2000, 2018; Hırçın Çoban & Sert, 2020; Mondada, 2014; Streeck, 2009).

There are many examples of behaviors in the literature that relate to the management of interaction in both L1 and L2 settings. Some of these are studies of individual behaviors (e.g., the head poke, Seo & Koshik, 2010), while others represent more complex gestalts (Mondada, 2014) of behavior. An exhaustive list would be impractical, but some of these interactional functions and their associated behaviors are listed

below:

- turn selection by using facial expressions and gestures (Streeck & Hartge, 1992) or pointing (Mondada, 2007; Nakatsuhara, 2011)
- the management of turn-taking through mutual gaze, averted gaze, shifts in gaze, and head movements (Goodwin, 1980; Greer & Potter, 2008; Rossano, 2012)
- maintaining extended turns in storytelling sequences through gaze direction and paralinguistic features (Tominaga, 2013)
- indications of turn completions through combinations of gestures and body motion with paralinguistic vocalizations (Keevallik, 2014), gesture holds (Groeber & Pochon-Berger, 2014), and gesture retraction (Mondada, 2006)
- the initiation of repair through gestures, and the mediation of meaning through raised eyebrows, mouth movements, and mutual gaze in mediation sequences (van Compernolle, 2013)
- indication of comprehension problems (trouble) through long silences, averted gaze, and smiles (Hırçın Çoban & Sert, 2020) and raised eyebrows and head displays (Oloff, 2018)
- repair initiations with mutual gaze and holding the floor for lexical retrieval with averted gaze (Burton, 2021a; Pekarek Doehler & Skogmyr Marian, 2022; Streeck, 2009)
- communicating trouble sequences and showing resolution through holds and their release (clusters of behavior held static during a period of time) (Burton, 2021a; Floyd et al., 2016; Oloff, 2018), head pokes (Seo & Koshik, 2010), and postural leaning forward (Rasmussen, 2014)
- complaining sequences through the use of paralinguistic features, gestures, facial expressions, eye gaze, and posture shifts (Skogmyr Marian, 2023)
- the maintenance of intersubjectivity in paired speaking tests through eyebrow flashes and gestures (Burch & Kley, 2020)
- communicating speech acts of greetings, farewells, and introductions with gestures such as waving

and haptics such as hugs and handshakes (Rylander et al., 2013)

These are just a few of the documented cases of how behaviors can contribute to interactional management in L1 and L2 settings. Because nonverbal behaviors and associated meanings do not have one to one relationships, however, exhaustive lists of behaviors and their associated possible meanings would not be meaningful, as the situational context provides the interpretative lens for these actions. Linguistic forms and their semantic intent also do not always have one to one relationships, but the possible range of associations is much more restricted. Mondada (2014) noted that

> participants might choose the way in which they format a particular action—and that these choices might vary, privileging either verbal resources, a combination of verbal/embodied resources or embodied resources alone… these choices might be constrained in interesting ways in situations of multiactivity, where participants distribute resources among various concurrent courses of action and often prioritize one over the other (p. 140)

Thus, there is no guarantee that any particular action or verbalization may occur during particular interactional-pragmatic sequences. Just as someone can as easily say "goodbye", wave, or both to a departing guest, speakers make choices depending on context and perhaps their own idiosyncratic preferences. Nonetheless, interactionally competent listener-receivers are able to interpret these various signals as particular courses of action.

Socio-affective information

One of the most powerful roles of nonverbal behavior is to convey affective information about emotions, attitudes, and stances (Kappas et al., 2013; Richmond & McCroskey, 2004; Singelis, 1994). These affective responses help to drive social interactions and may reveal information about underlying psychological states of speakers. Because of the importance of this topic, affect is treated in an entire subsection in this literature review following this one on nonverbal behavior. However, nonverbal behavior and affect are also an important drivers of social interaction, particularly the face, which has been described as "the primary site of affect displays" (Ekman & Friesen, 1969, p. 841). Social interactions are characterized by a wide range of both verbal and nonverbal behavior which conveys meaning about the

nature and status of the relationships amongst the speakers (Argyle, 1988; Mehrabian, 1972; Patterson, 1983). If one thinks of a group of friends catching up over coffee, the topic of the conversation will of course take center stage, but the visible reactions amongst the interactants will stand out as secondary sources of information. Smiling at a fellow group member may indicate deference towards that individual, or it may serve as a backchanneling device to acknowledge the other's utterance. A rise of the upper lip might appear as a sneer and could convey disdain or contempt for another member of the group, which could then lead to a topic-shift or a turn ending. While these behaviors are both quite noticeable, research has shown that even more subtle, almost imperceptible movements of the body, such as movements of the arms or the face, can also impact views of other individuals (Argyle, 1988; DePaulo & Friedman, 1998). These various behaviors may play an important role in communicating power dynamics such as dominance and submission (Tiedens & Fragale, 2003). In these cases, the information being conveyed can be both affective in its attitudinal and emotional components and social in its orientation towards organizing conversation.

Some socially oriented aspects of nonverbal behavior arise from individual differences in speakers. Culture (discussed in the next section) is an important moderator of such individual differences. Sex and gender, both interacting with culture, also appear to relate to social-forward nonverbal behavior. For example, there may be differences in the sensitivity of reading nonverbal cues, comfort with proximity and touch, and proportions of mutual gaze between men and women (Hall, 1984, 2006). Personality can also exert an effect on the production of nonverbal behavior, with extraverted, relaxed individuals showing higher signs of engagement such as closer proximity, greater proportions of mutual gaze, and higher overall expressiveness such as smiling (Neumann et al., 2009; Patterson, 1983; Patterson & Ritts, 1997). Even socioeconomic status may result in differences amongst speakers. Kraus and Keltner (2009), for example, found that speakers from more advantaged socioeconomic backgrounds conveyed a greater amount of disengagement (averted gaze and attention, doodling on paper), while speakers from more disadvantaged socioeconomic backgrounds generally conveyed a higher degree of engagement such as head nodding and laughter. In all of these cases, context and culture likely moderate the findings as well. Many other

individual differences have also been found to relate to differences in nonverbal behavior (e.g., Hall & Gunnery, 2013; Gifford, 2013; Nestler & Back, 2013, Rule & Alaei, 2016).

In interaction, speakers orient towards the behavior of their partner, often adjusting their behavior depending on the affective nature of the interaction. In equilibrium theory (Argyle & Dean, 1965; Patterson, 1973), speakers seek to maintain a balance between intimacy and behavioral expressiveness; behavior that violates norms of closeness, such as standing too close to a person, can be met with an opposing, equalizing behavior of moving away. Likewise, if one person maintains mutual gaze to the point of staring, the interactant may seek balance by looking away. In other cases, however, speakers may reciprocate the direction and valence of a behavior (Burgoon, 1978; Capella & Greene, 1982); that is to say, people may stand closer together, or a higher proportion of mutual gaze may lead to a corresponding level of mutual gaze in the speaking partner. In general, negative affect can cause a compensatory effect, and positive affect can cause reciprocation, but these naturally flow back and forth between speakers in complex parallel systems (Patterson, 2013). Affect does not drive all interactions, however. To some degree, behavior seen in social settings may be mimicked unconsciously (Chartrand & Lakin, 2013), and affect can be contagious. Behavior can also lead to various subconscious impressions made about speakers (Todorov, 2017), without any corresponding action. The issues of mimicry, contagiousness, and impression formation will be discussed in the following section on affect. Likewise, culture may be a strong determinant in how many of these socio-affective behaviors are conveyed and understood, and will be discussed next, as well as in the section on affect.

Cultural origins of nonverbal behavior

Anyone with experience talking to people from varying cultural backgrounds quickly realizes that there are differing norms when it comes to nonverbal behavior. In some contexts it is appropriate to shake someone's hand when greeting someone, while in others (particularly Japan), it is customary to bow with their hands by their sides. In the Mediterranean, it is often customary to kiss the check of an individual once, twice, or even three times, depending on the country or region that one is in. There are also differences between countries in the appropriacy of proxemics, or the distance one stands next to someone else, as well

as how long to hold mutual gaze. There are countless other examples. It is important to acknowledge the role of culture with nonverbal behavior in language testing settings because intercultural communication is often part and parcel of the experience. The test taker may exhibit cultural differences in nonverbal behavior that are systematically misinterpreted by raters from different cultural backgrounds and vice versa. While the previous sections detailed differences in the various types of information nonverbal behavior can provide, this section will briefly review some of the key issues to consider when dealing with intercultural encounters.

Culture provides the overall context for social encounters to occur, and it provides a framework for how social interaction should take place. Within the context of culture, Matsumoto and Hwang (2016) argued that

> [t]he function of communication is to allow for the sharing of social intentions, which facilitates social coordination. Cultural norms provide rules for the regulation of expressive behaviors, including nonverbal behaviors, to allow for the sharing of social intentions as part of communication… this underlying function of nonverbal communication vis-à-vis the function of culture is universal; the cultural norms and the manifestation of those norms in actual behavior, however, are different because of the various adaptations different groups have made to survive in their ecological contexts. (p. 77)

The authors make clear here that the semantic, cognitive, interactional, and affective roles of nonverbal behavior do not change (at least substantially) across cultures. What can change, however, are the nonverbal forms presented. Some forms do appear to be mostly universal, such as the use of verbal and nonverbal resources for repair (Dingemanse et al., 2015) and the tempo of turn-taking (Stivers et al., 2009), though learners do not always use these universal forms appropriately while learning a language (Pekarek Doehler & Pochon-Berger, 2015). Most forms, however, are not universal. The same function can be conveyed by different deployments of facial and bodily expressions in different cultures, and the same behavior may be seen in different functional sequences (Crivelli & Fridlund, 2018; Fridlund, 1994). This is also true within individual cultures, as detailed in the section on interaction (mutual gaze can indicate a turn release pattern

or also a repair initiation depending on context). Even regions within very similar cultural groups may develop slightly different muscle movements when conveying the same information depending on the setting (Elfenbein & Ambady, 2002). That is to say, there appears to be nonverbal "dialects" or "accents" that vary across cultural contexts in similar ways to language (Elfenbein et al., 2007; Marsh et al., 2003).

Researchers first began work in the domain of behavior and culture attempting to show that the nonverbal cues associated with affect displays were universal (Tomkins & McCarter, 1964). Building on their work, Ekman set out a research agenda to study the universality of affect judgements. In a series of studies, he and others found agreement between images of particular facial expressions and prototypical emotional categories in different cultural settings (Ekman, 1972; Ekman et al., 1969). These studies claimed to have found six universal emotional expressions: surprise, sadness, happiness, fear, disgust, and anger. Replications and meta-analyses have provided support to these findings (Matsumoto, 2001; Matsumoto et al., 2009). Nonetheless, these studies have faced criticism due to various methodological issues (Russell, 1994; Russell & Fehr, 1987), such as the use of Western faces and the types of scales used. In any case, these studies focused on a very narrow range of emotions rather than the much broader use of behavior in society.

Context, then, is critical to how cultural meaning develops. Different cultures may have different sets of display rules that determine whether particular behaviors should be displayed or not (Ekman & Friesen, 1969). For example, it may be appropriate to show anger in a service encounter in the United States, but a display of anger in a similar context in Asia would be far less appropriate. Cultures may also have different relationships between the public and private sphere, for example, and individuals may be more aware of their behavior in public (Matsumoto & Hwang, 2012). Some contexts carry a cultural weight with them that may be similar across international contexts, such as in a testing environment. Because testing environments, in particular those that are high stakes, are naturally imbued with uncertainty and anxiety for the test taker, behavior may change as a result of the setting. Guerin (1986) suggested that these types of contexts where the thoughts, intentions, and feelings of others (such as the examiner) are uncertain may cause individuals to restrict their behaviors and act more cautiously. Parkinson (2019), when discussing the

importance of context and culture, noted that "[t]he emotional meaning of faces may depend on the trajectories of action that they indicate or foreshadow and the centrality of those trajectories to our culturally specific prototypical representation of the emotion in question" (p. 89).

Cultural differences can sometimes lead to differential use in the forms of nonverbal behavior. As discussed earlier, emblematic gestures may differ substantially across cultural boundaries (Morris et al., 1979). Also discussed earlier are the differences in gestures that illustrate path movement (e.g., Kita & Özyürek, 2003), though other gestures forms may also differ depending on culture, such as counting movements (Pika et al., 2009). Gaze may also differ according to international background (Hall, 1963; Watson & Graves, 1966) or ethnic group (LaFrance & Mayo, 1976), such as when some groups may maintain mutual gaze longer than others. Paralinguistic cues, in particular prosody, speech rate, silence, and volume, can also convey differential emotional states or affective responses depending on culture (Sauter & Eimer, 2010; Sauter et al., 2010). For example, women in Japan may raise the pitch of their voice during telephone conversations to convey politeness, whereas Americans may be less likely to do so and instead use verbal mechanisms alone to convey politeness. There is relatively less research available on other behaviors such as eyebrow furrowing/raising, mouth movements (e.g., smiling), shoulder movements, head movements, and posture, though these behaviors may also exhibit certain differences around the world. More on how culture impacts nonverbal behavior through the lens of affect will be discussed in the section on affect below.

Nonverbal behavior and second language assessment

From this review so far, evidence has been presented that nonverbal behavior is a core aspect of communication that can convey semantic, cognitive, social-interactional, and affective information embedded in cultural contexts. When nonverbal behavior is seen and decoded by an interactant, either consciously or unconsciously, it becomes a central aspect of interpersonal perceptions. People see others and interpret their intended meaning from speech and the body, but they also infer information about what they might be thinking and feeling. Burgoon et al. (2016) argued that nonverbal behaviors have the capacity to shape and color our perceptions of speakers, even in the presence of contradictory verbal information.

Drawing on several decades of past research in the field of human communication, these authors distilled six tenets about the relationship between verbal and nonverbal communication:

1. On average, adults rely more on nonverbal cues than on verbal cues to determine social meaning.
2. Children rely more on verbal cues than adults do.
3. Adults rely more on nonverbal cues when verbal and nonverbal channels conflict than when these channels are congruent.
4. Channel reliance depends on the communication features at stake.
5. When the content in different channels is congruent, the meanings of the cues tend to be averaged together equally; when the content is incongruent, there is greater variability in how information is integrated.
6. Individuals have biases in their channel dependence. (pp. 221–224)

Burgoon et al. (2016) further stressed the co-occurring nature of nonverbal communication with verbal communication, as well as how the two interact:

> Because verbal and nonverbal signals arise simultaneously, the nonverbal channels can silently monitor the sender, send and receive feedback, express emotions, and define the interpersonal relationship all the while the verbal stream is conveying linguistic content. The nonverbal cues thus become the frame of reference against which verbal interpretations are checked. (p. 226)

Considering how nonverbal behavior impacts judgements, it is important to uncover the relationships between behavior and perceived L2 proficiency, as raters have found the visual realm salient when understanding certain verbal phenomena, such as breakdowns in fluency (Choi, 2022; Nakatsuhara et al., 2021a; Nambiar & Goon, 1993).

One theoretical and methodological framework that has been used when studying interpersonal judgements arising from nonverbal behavior is the Brunswik lens model (Brunswik, 1956; Nestler & Back, 2013; Hall et al., 2019), shown in Figure 2.3. This model accounts for the interpretation and accuracy of interpersonal perceptions between people. There are three core elements in the model that interact together:

a speaker's nonverbal behavior, their true underlying states or traits, and the judgement a listener makes about the individual based on impressions of their state or trait drawn from the visual realm. The relationship between the speaker's nonverbal behavior and their actual state or trait (measurable by asking the speaker or using validated physiological measurements such as galvanic skin response) is cue validity. In other words, this relationship defines to what extent a particular cognitive or affective element (e.g., excitement) is truly represented by nonverbal behavior (e.g., mouth agape, eyebrows raised, hands in air). The relationship between the true state or trait of the speaker and the listener's impression of that trait is interpersonal accuracy. In other words, if the speaker feels excited about a plane taking off, and the listener correctly interprets the speaker's emotion as excitement, the emotions were conveyed accurately; any misinterpretation results in inaccurate relationships and may suggest that particular affective responses are difficult to interpret accurately. Finally, of importance to this study is the relationship between the visible nonverbal behavior of the speaker and the way listeners interpret these when forming impressions, which is called cue utilization. In many non-laboratory settings, the speaker's actual trait or state is unknown, yet listeners (raters in this study) still use visible cues when making judgements. If, for example, particular facial cues result in ratings of excitement (and perhaps other tangential judgements), one can make inferences about how the cues impact certain perceptions. This model has been used to examine relationships between many interpersonal perceptions, such as intelligence (Borkenau et al., 2009; Reynolds & Gifford, 2001), self-esteem and expressiveness/warmth (Hirschmüller et al. 2018), personality and physical appearance (Naumann et al., 2009), and extraversion and likeability (Back et al. 2011; Borkenau et al., 2009).

Figure 2.3
Brunswik Model (Adapted From Hall et al., 2019)

In terms of communicative competence, there is strong evidence for, at a minimum, a supporting role of nonverbal behavior when conveying L2 ability through speech; in other words, listeners utilize the cues of nonverbal behavior when making at least some judgements about speakers, whether they are broadly holistic (competent in their L2) or specific to a subconstruct, such as interactional or strategic competence. Studies of rating mode showed that the visual domain introduces sources of variance in test scores (Choi, 2022; Gullberg, 1998; Nakatsuhara et al., 2021a; Nambiar & Goon, 1993), but necessary to this area of research is a documentation of how specific behaviors relate to language proficiency. A growing body of researchers, however, have made findings in this area using rater reports (Choi, 2022; Ducasse & Brown, 2009; May, 2009, 2011; Jenkins & Parra, 2003; Nakatsuhara et al., 2021a; Sato & McNamara, 2019; Thompson, 2016). The authors of these studies have determined that raters notice a range of different nonverbal behaviors, in particular the direction and holding of eye gaze, facial expressions, gestures, and posture. Another line of research has used post hoc analyses of video data to investigate differences in nonverbal behaviors for different proficiency profiles without rater reports (Gan & Davison, 2011; Neu, 1990). A final line of research has taken an empirical approach, documenting nonverbal behavior either

through raters, discourse analysis, or both to measure the quantitative impact of behavior on L2 proficiency outcomes (Gullberg, 1998; Kim et al., 2023; Thompson, 2016; Trofimovich et al., 2021; Tsunemoto et al., 2022). Overall, although raters primarily use verbal, linguistic information when making scoring decisions, raters also use nonverbal information when making decisions about L2 ability.

Rater reports. One of the most seminal papers considering the role of nonverbal behavior in ratings of L2 proficiency Jenkins and Parra (2003). The study was a qualitative analysis (discourse and video analysis) of eight Spanish and Chinese-speaking international teaching assistants taking part in a paired-format L2 speaking test. The authors analyzed written comments on the eight test takers' performances left by their individual examiners, as well as think-aloud protocols on all performances from one rater. They found that several nonverbal features contributed positively to the raters' perceptions of communicative effectiveness. These included nonverbal backchannels/receipt tokens (e.g., head nodding), displays of affect (e.g., laughing, smiling, frowning), mutual gaze, forward body leaning, and paralinguistic features. Importantly, showing engagement and listening comprehension (maintaining eye contact, leaning forward, and backchanneling) was an important feature that raters took into account positively when test questions were asked. Features that negatively affected perceptions included extended silence, an overall lack of expressiveness (e.g., an absence of head nodding, stiff posture, gaze aversion), a lack of eye contact, a stiff posture, and non-target-like prosody and tone. Another key takeaway of this study was that test takers with borderline-passing linguistic skills were able to compensate for their weaknesses by taking an affective, actively communicative stance, which boosted their scores in comparison to other borderline speakers, while a rigid, inexpressive stance negatively impacted test takers' scores.. "By making use of nonverbal features of conversational interaction, [test takers] can convince the evaluators that they have a higher level of proficiency than may in fact be the case from a purely linguistic perspective" (Jenkins & Parra, 2003, p. 100).

Three other studies discussed nonverbal in regard to interactional communication (Ducasse & Brown, 2009; May, 2009, 2011), using raters to uncover elements that factored into the raters' judgements. Ducasse and Brown's (2009) raters oriented strongly towards nonverbal behavior as contributing to

interactional competence, contributing to "the presence or lack of a 'physical' nonverbal fluency" (Ducasse & Brown, 2009, p. 433). Raters commented that mutual gaze was a positive attribute, while averted gaze was a negative attribute, without any reference to the conversational context of the gaze behavior. Gesture also contributed to positive and negative judgements; positive with co-speech gestures that illustrate affect, and negative when gesticulated and overly frequent. Similar to Jenkins and Parra (2003), nonverbal backchannels were seen as evidence of attention and listening comprehension and perceived positively. May (2009) analyzed four raters' comments about 12 test taker's interactions, considering elements contributing to co-constructed interaction. She also found evidence of nonverbal behavior contributing to successful communication. Closed-off body language, such as averted gaze and relative inexpressiveness, was perceived negatively, while establishing eye contact, gesturing appropriately, and nodding helped to sustain successful interaction. May (2011) extended this study and again found strong evidence for the role of nonverbal behavior in interactional competence. Notably, behaviors showing a desire to communicate were perceived as positive (maintaining mutual gaze, expressiveness, nodding, and using gestures) while behavior showing disinterest (avoiding eye contact, inexpressiveness, relatively rigid posture, facing away from interactant) were perceived as negative. She mentioned that:

> body language, although not mentioned in the rating scale, featured so prominently in the raters' evaluation of interactional effectiveness at all stages of the rating process. Raters' perception of body language… is also linked to assertiveness through communication, working together cooperatively, and contributing to authentic discussion. (May, 2011, p. 137)

May (2011) made a strong case for the interpretation of behavior through affect, which will be described in more detail later.

In a similar format, Sato and McNamara (2019) also considered factors that led to impressions of communicative success, but they used untrained, novice raters instead of operational raters. As described earlier, the authors found that a range of nonlinguistic criteria were related to communicative competence, of which nonverbal behavior was a small but important category, making up about 10% of the comments. Similar to the previous studies, co-speech gestures were seen positively, while gesticulation through self-

adaptors (e.g., playing with a ring) was seen as irrelevant to the context, appeared to indicate anxiety, and were seen negatively. Eye gaze was also an important behavior for these raters as it signaled various affective stances. Mutual gaze again was seen positively, and averted gaze was seen negatively without exception. As with May (2011), Sato and McNamara (2019) found that nonverbal behavior and affect were closely linked in the raters' judgements, as raters oriented to the socio-affective meaning-oriented output of the test takers' nonverbal behavior.

Two additional studies (Choi, 2022; Nakatsuhara et al., 2021a) primarily considered the impact of rating mode (video vs audio-only) and were discussed previously. However, these authors also elicited rater reports to investigate how raters perceived performance under the different rating conditions. Nakatsuhara et al. (2021a) found that a combination of mouth, eye, hand, and body movements helped enhance comprehensibility when pronunciation and fluency were less controlled. Behaviors also led to a clearer understanding of the test takers' affective stances, such as engagement, confidence, and the desire to communicate. Importantly, their raters mentioned that during disfluencies, nonverbal behavior provides evidence of whether breakdowns are due to a lack of comprehension, inadequate linguistic resources, or considerations of content. Understanding these distinctions was critical for raters when assigning appropriate fluency scores, as the scale categories included rationale for breakdowns. Raters also mentioned how inaccuracies were less noticeable when nonverbal behavior was visible. Overall, the authors remarked that video-based rating gave examiners a more complete view of the test taker's communicative competence, which therefore made them more confident in the scores they awarded.

Choi (2022), in a similar type of analysis, considered the verbal reports of eight raters when rating in three modes: audio-only, video with the examiner visible, and video without the examiner visible. She also found that video-based rating was more informative than audio-only rating, as raters had a fuller picture of test takers' performances. "Facial expressions, eye-gaze, and head orientation were commonly mentioned nonverbal cues that affected raters' perception towards test takers, signaled test takers' struggle during their responses, and showed test takers' focus during responses" (Choi, 2022, p. 151). Averting gaze by looking down, for example, was mentioned as evidence of problems with linguistic access, while

problems accessing content were shown through shifting gaze back and forth. Similar to Nakatsuhara et al. (2021a), nonverbal behavior provided important information about test takers' fluency, as seeing behavior allowed them to make more informed judgements. Also similar to Nakatsuhara et al. (2021a) was that nonverbal behavior, in particular gaze and head orientation, informed the raters of affective traits, namely engagement, desire to communicate, and confidence. Mutual gaze was preferrable for the raters, while averted gaze was seen negatively. An interesting finding in this study was that some raters found the visible rating format to be distracting, as processing nonverbal behavior took away focus on purely linguistic features. This may have been because the raters were used to rating formats in which accuracy was emphasized, which also links back to Nakatsuhara et al.'s (2021a) finding that inaccuracies were less noticeable in the video format.

Behavior at proficiency levels. Neu (1990) investigated nonverbal behavior in relation to communicative success. Drawing from the literature at the time, she presumed that important to communicative success were:

1) the appropriate use of meaningful gestures
2) nonverbal gestures in aid of speech difficulties
3) nonverbal gestures appropriately synchronized with the verbal channel
4) the degree of synchrony between interactants (Neu, 1990, p. 122)

She conducted a case study of two individuals taking a placement test for international students, choosing these two based on their similar scores yet diverging performances. She analyzed the performance data using the Foster system of discourse analysis (described later), transcribing verbalizations, facial/head movements, gestures/arm movements, and posture. She found that after discourse analysis, one test taker, Yama, had been rated lower than expected. She hypothesized that his lower ratings may have been due to his relatively inexpressive stance, with few facial expressions, shifts in posture, or nonverbal backchanneling. Furthermore, this test taker used gestures that were not aligned with speech, were complex, and highly frequent, which gave the impression of struggling. The second test taker, Ahmed, appeared to have inflated scores when compared with his actual discourse. Neu noted that his body posture was much

more relaxed and dynamic than Yama's, giving an air of confidence. He also used eyebrow raising, foot tapping, nodding as an interactional device, and tilted his head frequently to show engagement, to initiate repair, and to function as gestural beats indicating semantic units of speech. His gestures co-occurred with speech and were overall simpler than Yama's.

> Ahmed… appears to take control of the process…[his nonverbal behaviors] allow Ahmed to compensate for his weak verbal skills. When Ahmed cannot understand something, he bluffs his way through… by being nonverbally strategically competent in conversational interaction, Ahmed gives the impression of being more verbally communicatively competent than he is. Because Yama has not acquired nonverbal strategic competence in conversation, his verbal channel is perceived as less fluent than it really is. (Neu, 1990, p.136)

Thus, similar to Jenkins and Parra (2003), Neu (1990) largely determined that showing engagement, expressiveness, and confidence through nonverbal behavior were critical elements that helped Ahmed, and the lack of such expressions brought Yama's scores down. Their nonverbal behavior became critical to their unfolding conversations given its role in managing interactional moves of topic initiation and turn taking.

Gan and Davison (2011) used multimodal conversation analysis to document gestures that test takers used in a group-based interactive speaking test in Hong Kong. The analysis consisted of interactions within two groups characterized by different score profiles: low and high. Using McNeill's (1992) categorization system, they coded iconic, metaphoric, beat, and deictic gestures, but not self-adaptors or emblems. They found that the higher scoring group used co-speech gestures to provide detailed lexical meaning to utterances, emphasize particular ideas and suggestions, and aid in interactional management. Although not the focus of the study, the test takers also integrated eye contact, facial expressions, and body posture into their interactional moves. The lower scoring group displayed markedly different nonverbal behaviors in their group interaction. Some interactants were fairly rigid with very few visible nonverbal behaviors, while others used gestures that did not align with speech and were self-adapting in nature (e.g., scratching hair). These unsynchronized gestures indicated problems accessing language and aligned with breakdowns in fluency. These learners, however, did use deictic gestures (i.e., pointing) to assign turns as

a form of interactional management. The authors concluded that "synchrony of gestures with speech and other nonverbal acts proves interactionally, linguistically, and cognitively challenging, and their gestures seemed to be utilized predominately at the paranarrative level and to be involved in self-organization processes" (p. 116).

Quantitative studies. Gullberg (1998) is one of the first researchers who used a quantitative analysis to study the impact of nonverbal behavior—limited to gestures—on ratings of overall proficiency. She had a relatively small sample of 20 raters score 20 narrative speech samples in French and Swedish (10 raters and 10 samples per language). In each language set, half of the samples were L1 speakers of the language and the other half were L2 speakers. The raters scored the samples on a 5-point Likert scale of overall linguistic level with descriptions of only the endpoints (1 was the lowest, 5 was the highest). She found that the number of perceived gestures (observed by the raters) was the only variable that correlated with proficiency evaluations, which correlated quite strongly at .75. In other words, the more that raters noticed gestures, the higher they rated the test taker's L2 proficiency. The actual number of most gestures produced (tallied by Gullberg) did not correlate with proficiency except for iconic gestures, which related to meaning making. An additional interesting finding was that raters' perceptions of gestures were marked by error in comparison to the actual gestures produced. Raters only noticed some of the gestures, under- and overestimating the amount of gesturing in different speakers. "What constitutes 'many' gestures must therefore be assumed to reflect qualitative differences between gestures, or gesture types, with respect to how perceptually *salient* they are, in a broad sense" (Gullberg, 1998, p. 203).

In a small-scale, mixed methods study, Thompson (2016) analyzed the relationships between the frequency of eye contact, gestures, and smiles with holistic score outcomes using the IELTS rating rubric. She also considered modality differences between audio-only and audiovisual rating. She followed up with raters by conducting stimulated verbal recall sessions. She recruited four Canadian raters to conduct four practice oral proficiency tests with four mainland Chinese test takers. She found that some behaviors appeared to be associated with score outcomes. For example, a higher rate of self-adaptors (e.g., head scratching), a greater amount of mutual gaze, and non-Duchenne smiles (less authentic smiles where the

muscles around the eyes do not move) were associated with lower scores. In at least one test taker, more representational gestures and authentic, Duchenne smiles may have compensated for limited lexical resources to enhance scores. Additionally, ratings on fluency and pronunciation were found to be consistently higher when rated with visual information present in the audiovisual mode, though broader differences in modality were inconclusive. Although the sample was quite small and findings are largely inconclusive, the study is notable as it is one of the only studies to consider the impact of specific behaviors on discrete proficiency outcomes.

Tsunemoto et al. (2022) considered the relationships between various facial behaviors and gestures and ratings of comprehensibility, accentedness, and fluency. They had 60 novice raters assess videos of 20 L2 English speakers with Chinese and Spanish as their L1 narrating a story in English. Ratings were conducted in three audiovisual conditions: audio with a static image, audio with only facial behaviors visible, and audio with both face and body visible. For behaviors, they tallied raw frequency counts of head movements (tilts, shakes, and nods), eyebrow movements (raises and frowns), averted gaze (looking away, up, aside, and down), blinking, smiling, laughing, pursed lips, referential gestures (iconic, metaphoric, and deictic), and beat gestures. Access to the face and the face and body was associated with progressively higher comprehensibility ratings. That is to say, comprehensibility scores were higher with a visible face than a static face, and the body and face scores were higher than the face alone. Lower accentedness scores, however, only occurred when the face and body were both visible. There were no differences in fluency scores as they related to viewing condition, in contrast to other studies on modality differences (e.g., Choi, 2022; Nakatsuhara et al., 2021). They additionally found no significant differences in behavior production as a result of cultural background, and furthermore cultural background did not interact with scores awarded. Some specific behaviors correlated with speech scores. They found that more frequent eyebrow movements (both raises and frowns) related to less accented and more fluent speech. They speculated that eyebrow movements may have enhanced prosodic information available to the viewers, signaling content and indicating phrase boundaries, possibly in combination with hand gestures. Averted gaze associated with higher comprehensibility. They speculated that gaze aversion may have helped speakers in their cognitive

processing, leading to enhanced performance. They further considered the possibility that gaze aversion "was expected by an external observer, with the consequence that this visual behavior alleviated at least some processing burden for the rater" (Tsunemoto et al., 2022, p. 678). Other behaviors, including behaviors indicating positive emotions, were not related to speech ratings.

Using data from the same corpus as Tsunemoto et al. (2022), Kim et al. (2023) selected a different set of 40 L2 participants taking part in paired interaction rather than monologue narration. The participants were from a diverse range of L1 backgrounds. The L2 participants recorded perceived fluency scores about their conversational partner on a 100-point sliding scale with disfluent and fluent as endpoints. The authors extracted raw frequency counts of the incidence of the same nonverbal behaviors as Tsunemoto et al. (2022): head movements, eye gaze direction, eyebrow and mouth movements, and gestures (representational and beats). They found positive relationships between eyebrow and mouth movements (smiling) with fluency (.31 and .34, respectively), but a negative relationship between non-beat representational gestures and fluency. They speculated, similar to Tsunemoto et al. (2022), that eyebrow movements served to highlight speech prosody, which tied into judgements of fluency. They further mentioned that smiling led listeners to judge their conversational partners as less anxious and more engaged, which served as a proxy for fluency measures. Regarding representational gestures, Kim et al. (2023) speculated that gestures broke the continuity of speech, shifting conversational partners' attention away from the interaction and making the speaker appear less fluent. However, these findings are contradictory to broad agreement in the literature that co-speech occurring representational gestures can serve to facilitate speech processing for listeners (Drijvers & Özyürek, 2017; Jenkins & Parra, 2003; Kelly et al., 2008; Gullberg, 1998, 2006).

Similar to the previous two studies, Trofimovich et al. (2021) investigated the impact of various measurements of engagement on comprehensibility in the same corpus as Tsunemoto et al. (2022) and Kim et al. (2023). Amongst these measures were nonverbal backchanneling (nodding) and measures of positive affect. They counted the raw frequency of nods, smiles, and verbalized mentions of positive affect in 36 dyads from various national and linguistic backgrounds. Comprehensibility was measured by each partner in the dyads. The authors found that comprehensibility was predicted by a greater number of nods and the

use of encouraging verbal behavior. Positive affective behaviors were not a significant predictor in the model, but the authors noted these behaviors occurred infrequently. Nonetheless, they did correlate positively with measures of comprehensibility.

Summary. To summarize the findings from studies relating nonverbal behavior to studies of L2 proficiency, Table 2.1 lists behaviors, their effects, and the sources documenting those effects. It can be seen that mutual gaze is generally regarded positively, as it showed engagement and confidence, a desire to communicate with the examiner, and occasionally listening comprehension. Averted gaze is almost always seen as negative, as it indicated some sort of struggle or disengagement. Similarly, expressive behavior (eyebrow movements, head movements, smiles, and other less defined body movements) is almost always seen as a positive attribute. These may indicate interactiveness and engagement but could also highlight prosodic information that may have likewise indicated greater spoken fluency. Any lack of such (including extended silence) may be regarded as negative, especially if an inexpressive stance characterizes the interaction. Gesture depends on where the gesture occurs and its frequency. Co-speech gestures are almost always seen positively as they add important lexicosyntactic information to speech and help test takers to manage interactions. Gestures occurring during silences or occurring too frequently are perceived negatively due to their compensatory role during breakdowns. This is frequently the case with self-adaptors, which are seen as a coping mechanism. Importantly, many students observed that nonverbal behaviors convey affective information, such as confidence and engagement, that raters use when perceiving different aspects of language proficiency.

Table 2.1
Summary of Nonverbal Effects on L2 Perceived Proficiency

Behavior	Impact on L2 proficiency	Source
Gaze		
Mutual	+Engagement, desire to communicate, confidence, and listening comprehension – Could associate with lower scores (Thompson, 2016)	Choi (2022), Ducasse & Brown (2009), Jenkins & Parra (2003), May (2009, 2011), Nakatsuhara et al. (2021a), Sato & McNamara (2019)
Averted	– Negative impact, indicating anxiety, access of language or content + Can enhance comprehensibility (Tsunemoto et al., 2022)	Choi (2022), Ducasse & Brown (2009), Jenkins & Parra (2003), May (2009, 2011), Nakatsuhara (2021a), Sato & McNamara (2019)
Mouth: Smiling	+ Enhanced fluency, raised engagement, lowered perception of anxiety + Possible relationship with comprehensibility + if Duchenne (authentic) – if non-Duchenne (inauthentic)	Kim et al. (2023) Thompson (2016) Trofimovich et al. (2021)
Eyebrows		
Raised	+ Interaction management + Lowers accentedness + Raises fluency due to prosodic cues	Kim et al. (2023), Neu (1990), Tsunemoto et al. (2022)
Frowning	+ Lowers accentedness + Raises fluency due to prosodic cues	Kim et al. (2023), Tsunemoto et al. (2022)
Head		
Nods	+Engagement, listening comprehension, and interaction management + Comprehensibility – Negative impact if missing	Jenkins & Parra (2003), May (2009, 2011), Neu (1990), Trofimovich et al. (2021)
Tilts	+ Engagement and interaction management	Neu (1990)
Posture		
Leaning forward	+ Engagement and listening comprehension	Jenkins & Parra (2003)
Leaning backward	+ Confidence, low anxiety	Neu (1990)
Rigidity	– Negative impact	Gan & Davison (2011), May (2009, 2011), Neu (1990)
Gesture		
Representational (iconic, metaphoric, and deictic)	+ Co-occurring with speech, added lexicosyntactic information + Interaction management (deictics) – In silence, gesticulated, or too frequent – Lowers fluency, breaks attention (Kim et al., 2023)	Ducasse & Brown (2009), Gan & Davison (2011), Gullberg (1998), May (2009, 2011), Neu (1990), Sato & McNamara (2019), Thompson (2016)
Beats	+ Emphasizing semantic units + Interaction management	Gan & Davison (2011), Neu (1990)
Self-adaptors	– Indicated anxiety, problems with lexical access, and breakdowns in fluency	Gan & Davison (2011), Sato & McNamara (2019), Thompson (2016)
Other		
Silence	– Negative impression if extended	Jenkins & Parra (2003)
Expressiveness	+ Positive impact	May (2011), Jenkins & Parra (2003), Neu (1990)
Inexpressiveness	– Negative impact	Gan & Davison (2011), Jenkins & Parra (2003), May (2009, 2011), Neu (1990)

As the various authors noted, the findings generally align with those from the broader literature on nonverbal behavior in that behavior can convey cognitive, semantic, social-interactional, and affective information. When L2 speakers use these behaviors in target-like ways that show engagement, confidence, and a desire to communicate, they are seen positively. When behaviors show anxiety, do not align with speech, occur in non-target-like ways (being too frequent), or are generally rigid and not present, L2 speakers are perceived as less proficient. Raters use behavior to infer sociocognitive information about speakers—language proficiency—often through an affective lens. It is also possible that raters are impacted by affective responses, altering their own emotional stance towards the interview and coloring their judgements about the test takers. Because of the importance of the ability of nonverbal behavior to convey affect, this review now turns to that topic in more detail.

Affect

Analogous to the form-function relationship between language and semantics, nonverbal behavior (a set of forms) conveys information to listener-interactants. The previous section covered a sample of the cognitive, affect, and interactional information that nonverbal behavior conveys. Affect, however, has received substantial attention in the literature due to its close connection with body language, and because of its importance in human communication, this review will now cover this topic in more detail. Affect may include mood, emotions, interpersonal attitudes, and other personality states or traits. It is present alongside the semantic content of utterances in all types of communication. We infer stances and feelings from written texts based on the choice of words and discourse formulation on the page, but in speech we are able to draw from the repertoire of both language and nonverbal behavior to make inferences about thoughts, feelings, dispositions, and motivations of others. These inferences and attributions may then be used when evaluating others in interpersonal encounters. The previous section has discussed some of the forms and functions of nonverbal behavior, as well as how they can impact the evaluation of language proficiency. In practice, individuals may be less attuned to the specific behaviors that they see and instead formulate evaluations based on the affective responses they infer from their interactants. In this section, I begin by defining key terms, followed by a discussion of the cognitive, social, and cultural origins of affect. I will describe how

affective responses relate to language achievement, how they impact interpersonal relationships, and how affect has appeared in the language testing literature in relation to evaluations of proficiency.

Definitions

In general language use, the terms emotion, mood, and affect are often grouped together as the same concept. This is also true in the academic literature in theoretical and empirical studies (Briner & Kiefer, 2005). These terms are often grouped together as affect (Frijda, 1994; Scovel, 1978) or emotion (Pavlenko, 2006). In applied linguistics research, Pavlenko (2006) lamented that emotions have often been reduced to "a laundry list of decontextualized and oftentimes poorly defined sociopsychological constructs, such as attitudes, motivation, anxiety, self-esteem, empathy, risk-taking, and tolerance of ambiguity" (p. 34). Though certainly similar, emotions, mood, and affect are distinct. Each is rooted in varying neurophysiological processes and psychosocial phenomena. Here I provide some definitions of each, though noting the caveat that disagreements on these definitions persist (Russell, 2012).

Emotion. Scherer (2005) characterized emotions as a) being a reaction to some sort of stimulus, b) varying in intensity, but often more intense than moods, c) lasting a relatively short period of time, and d) resulting in some change in behavior. Emotions may consist of cognitive, physiological, motivational, expressive, and experiential factors (Scherer, 2005). Cognitive factors refer to an assessment of the stimulus event; physiological factors are the body's internal changes due to hormonal release and reactions in the nervous system; expressive factors comprise all of the automatic, unconscious behavioral embodiment of the emotion in the body and face; motivational factors are the actions resulting from the emotion; and experiential factors refer to the subjective, personal experiences of an emotion by the individual (Lischetzke & Eid, 2003; Scherer, 2005). Experiential factors are more commonly known as feelings, which may not align exactly with the internal physiological factors at play in emotions, as individuals may misinterpret or be unwilling to communicate their true emotions (Tran, 2007). Plutchik (2001) proposed eight core emotions: joy, surprise, anticipation, trust, sadness, fear, anger, and disgust.

Mood. Affective states that are generally lower in intensity than emotions, rather diffuse in nature, longer lasting than emotions, and often without a definite stimulus event or focus are called moods

(Frijda, 1994; Lochner, 2016; Tran, 2007). Moods are also more global in nature and may be comprised of unconscious background sensations (Lischetzke & Eid, 2003). Mood, unlike emotion, is purely subjective and an individual's perception of their internal state (Lochner, 2016). Similar to emotion, mood may also be due to physiological changes and result in behavioral shifts in the body and face (Lochner, 2016). Some examples of mood may be general contentment, ennui, or depression.

Affect. Although affect commonly includes emotion and mood (Frijda, 1994; Scovel, 1978), it also refers to a range of characteristics aligned more closely with personality. These characteristics may include broadly pleasant or unpleasant experiences (Frijda, 1994), dimensions of personality traits or states (Diener et al., 1995; Watson et al., 1988), or attitudes (Scherer, 2005). Similar to emotion and mood, affect may also be rooted in physiological mechanisms and result in changes in bodily behavior. Affect is distinguished from emotion and mood in that while emotion and mood are generally internally realized, affect is something *observed* and *perceived* by others as a trait or state. An individual may convey a particular personality characteristic or attitude at one particular moment (state), or they may be disposed in their reactions to experiences that appear as trait elements of their personality (Lischetzke & Eid, 2011). Affect may thus be transitory or even stable for long periods of time, perhaps even a lifetime (Mehrabian, 1996). Trait affect may thus serve as a lens to moderate emotions or moods people experience in certain situations (Lischetzke & Eid, 2003). Examples of affect common in SLA literature are anxiety, motivation, and willingness to communicate.

In this dissertation, the focus is on the perceived, outward-facing emotions, feelings, orientations, and stances of what individuals perceive, and for this reason the focus will be on affect. Affect best captures the range of perceptions a listener may have of another person, such as being happy, confident, or engaged. I will follow trends in psychology and group the various concepts together as affective phenomena (Tran, 2007), and I will refer to them broadly as affect. When discussing a particular instance of visible affect, I will use terms such as affect displays, affective responses, or affective reactions. What is important in the context of this study is not the actual emotion or mood the individual experiences (an internal state), but rather how others perceive and decode the nonverbal behavior from a second language speaker as an

external social experience.

Cognitive origins of affect

People are continuously observing and reacting to their environment in our everyday lives. In many of these moments, people experience feelings about what they see, and they may choose to express those feelings for a number of reasons. Perhaps because every human experiences these internal reactions to stimuli, emotions and attitudes have historically been characterized as largely internal cognitive phenomena. A person may feel happy or proud because they won an award, and thus project an outward affective expression of being focused, attentive, and confident. Another person may feel sad, angry, or shameful because they failed a test, and they may be seen as anxious, disengaged, or non-interactive. People often attribute these affective reactions as living purely within the individual.

Research in psychology has historically supported this supposition. William James (1884) defined emotions as personal, subjective experiences that arise from an individual's internal senses and observations of the world around them. Supposing James is right, the body should have consistent reactions to particular phenomena, with the body and brain showing particular patterns when feeling sad, angry, or happy. Attempts to identify internal bodily patterns associated with emotions have, however, come up short. Although there are some patterns that differentiate affective responses (Kreibig, 2010)—such as rising blood pressure's association with anger rather than happiness—there are no consistent patterns in the autonomic nervous system that characterize each particular emotion (Siegel et al., 2018).

Schachter (1964) developed a two-factory theory to compensate for the lack of a one-to-one relationship between autonomic response and emotions. In this theory, the brain manages physiological arousal in light of situational contexts, thus producing emotional responses that take both into account. In this theory, heightened arousal when receiving good news would be characterized as happiness, while heightened arousal when being slapped would be felt as anger. Context, in other words, retroactively translates the arousal of the emotional experience as valence. Although validation studies of this theory have failed to replicate the initial findings (e.g., Manstead & Wagner, 1981), the implications are important because it recognizes that outside stimuli are the driving force behind the valence of affective responses.

Barrett (2017) extended the nuanced view of outside stimuli interacting with internal changes in the body, showing different patterns of bodily responses characterizing particular expressions of an emotion in different contexts. Her theory moves away from claiming that the body produces a specific response, but it maintains that emotions are subjective experiences with externally classifiable representations.

Other researchers have placed an even greater emphasis on the role of situational contexts in their theories. In appraisal theory (Arnold, 1960; Lazarus, 1991), it is the evaluation of particular environmental stimuli that activates emotions. This theory argues that emotions rarely appear spontaneously; they are physiological responses to objects, entities, and actions being witnessed in the world. Emotions, then, are feelings happening inside the body and also a reaction to emotionally primed situations (Frijda, 2005). The "why" behind an emotion plays just as much of a role as the emotion itself; the award the student won is what drives the happiness to begin with. Different emotions in appraisal theory are driven by dimensions of motivational relevance, motivational congruence, and accountability (Smith & Lazarus, 1993). In terms of relevance, situations that are not deemed as impactful to an individual do not provoke affective responses. Incongruent events hinder our progress and cause negative emotions, while congruent events help to provoke positive feelings. Finally, accountability determines whether the source of the event is oneself or external. Thus, happiness about an award would be relevant, congruent, and externally caused, while the source of pride would be similar but internal instead of external. Nonetheless, critiques of this model have argued against this type of strict causality, leading reformulations that show that appraisals can be consequences of affective responses as well (Lazarus, 1991). Events are thus appraised in relation to the social context at hand; it is not the award itself that causes happiness, but the appraisal of having earned it after working hard for a period of time.

The lack of characterizable physiological responses and the growing importance of context has led some psychologists to argue that emotions may be more complex than previously thought, occurring instead as social phenomena, and arising between individuals. Emotions may then be social, distributed amongst individuals, and thus the result of interpersonal interactions (Parkinson, 1996). Happiness may not only arise just because an individual won an award after working hard on a particular project. It may also be the

result of the person's knowledge of their parents' and teachers' expectations and resulting pride in response to the student's award. I turn to the social origin of affect in the next section.

Social origins of affect

The attribution of affect to social settings is intuitive. People are usually able to claim a source to their emotions, such as when children claim that "so-and-so made me mad!" These affective responses may serve to prepare people physically for further action, resulting in certain emotional dispositions (Arnold, 1960). The angry child in this example would then be prepared to take action or flee from the object of their anger. Frijda (1986) made the explicit case that affect serves as a relational device between individuals "that establishes or modifies a relationship between the subject and some object or the environment at large" (p. 13). Some affective responses may then be dependent on social relationships and originate in social encounters, which led Mesquita (2022) to claim that "emotions *are* for acting, and particular for acting in the social world" (p. 53, emphasis in the original).

In one socio-affective model, de Riviera (1977) argued that emotions can be distinguished by the type of activities occurring between individuals rather than within them. These emotions are thus colored and contextualized by the social relationship itself instead of the physiological responses or the objects and people being appraised. An emotion such as happiness, then, serves to bring people together, and a feeling such as anger would serve to eliminate a threat or challenge. One can imagine a complex attitudinal response such as confidence serving dual roles as an internal appraisal of one's abilities and also an affective stance that communicates to others that the individual is ready, willing, and able to tackle a challenge. This push and pull between the individual and social directions of affective responses may be an inherent characteristic of all emotions (de Riviera & Grinkis, 1986). According to Parkinson (2019), "emotion's special ingredient is its capacity to align and realign people's relations with each other and with objects and event in the shared environment" (p. 2).

In these examples, the emotions underlying affect are still to some extent residing in the individual's responses to their environment. Some have argued, on the contrary, that while this may be the case for some sets of emotions, others originate between individuals (Boiger & Mesquita, 2012; Frijda &

Mesquita, 1994; Rimé et al., 1991). Mesquita (2022) differentiated between these two types as MINE emotions (mental, inside the person, and essentialist) and OURS emotions (outside the person, relational, and situated) (pp. 23–24). MINE emotions are those that we traditionally associate with, such as feeling happy when winning an award. OURS emotions depend on the shared experiences between others. Uchida et al. (2009) provided evidence of OURS-based emotions of happiness in the ways that Japanese and Americans talked about feelings. Japanese individuals were much more likely to talk about the feelings of a group as a whole, and Americans were more likely to talk about their individual emotions. Japanese participants were also more likely to identify happiness in photographed individuals only when they were surrounded by smiling people, whereas Americans identified happiness in the individual regardless of the facial expressions of their surrounding group. Other findings have corroborated these claims (Masuda et al., 2008, 2012).

These findings point to evidence that emotions and the broader realm of affective phenomena are socially constructed. In L2 testing settings, there is some evidence that examiner behaviors exhibiting positive or negative affect can impact the corresponding affective responses of test takers as well (Briegel-Jones, 2014; Plough & Bogart, 2008), such as a test taker feeling less anxious with a friendly examiner. However, as apparent in the previously cited studies, there may also be important differences in how emotions are conveyed and decoded in different cultural settings. Given that L2 testing contexts are almost always intercultural encounters as well, the next section details some of the ways culture plays an important role in our understanding of affect.

Cultural origins of affect

Emotions and affect have been discussed as having origins that are simultaneously cognitive and social. Related to the social phenomenon of emotions is the outstanding question of whether emotions are universal across all cultures. Ekman's (1972) neurocultural theory argued that certain basic emotions (e.g., happiness, sadness, anger, fear, disgust, and surprise) are hard-wired into the human nervous system and shared across all cultures. Furthermore, he argued that those emotions had discrete representations in muscle movements in the face. Any deviation from these affective facial behaviors and their underlying

emotions was due to social learning or display rules—cultural norms that dictate the appropriacy of expression of certain emotions in social contexts. Ekman's theory was supported by findings that allegedly showed that diverse cultures were able to accurately identify—at a rate greater than random chance—particular emotions in connection with images of Western faces (Eckman et al., 1969). Eckman et al. argued that this provided evidence of some degree of universality in at least some emotions, thus rejecting cultural relativism.

In the section on nonverbal behavior, studies were presented that argued that Ekman's link between emotions and behavior were consistent in some ways across cultures. Not all psychologists agree, however, that better-than-chance identification of emotional displays means that emotions are universal. Some reject the premise of emotional universality, arguing that cultures may differ in how emotions are *felt* (Mesquita, 2022), how they are encoded and decoded (Crivelli et al., 2016; Elfenbein & Ambady, 2002), and how they describe emotions through their internal lexicons of emotional vocabulary (Wierzbicka, 1992). For example, in a study across 2,500 languages, Jackson et al. (2019) found low semantic similarity among 24 emotion concepts (English-based concepts) across cultures. They found that the only term that was similar across nearly all cultures was broadly "feeling good." Other terms, such as bad, love, happy, and fear, occurred in as many as 70% of the languages to as low as 15%. Some languages may lack familiar emotional terms, such as "sadness" in Tahitian (Levy, 1973). Other languages have emotional terms that have no direct translation in English, such as "amae"—a feeling of coziness and tenderness when being quasi-dependent on a parent—in Japanese (Morsbach & Tyler, 1986).

Different cultures may experience and interpret emotions somewhat differently as well. For example, conveying happiness does not appear to be perceived equally across all cultures. In the American cultural tradition, happiness is an integral part of social interactions (Wierzbicka, 1994) and may convey success, achievement, pride, superiority, and self-esteem (Uchida & Kitayama, 2009; Kitayama, et al., 2006; Shaver, et al., 1987). Happiness is thus seen as positive and desirable, and individuals showing happiness may benefit when dealing with others, solving problems, and negotiating outcomes. This is not the case in all cultural contexts. For example, in one study Japanese students differed from their American counterparts

in that they associated happiness with both positive and negative traits; happiness was seen as temporary and fleeting, but also potentially disruptive to social groups because of its potential to cause envy (Uchida & Kitayama, 2009). Differences in "the antecedents, the actions, the reactions from others, the consequences, and arguably the associated feelings" of emotions can be found in nearly every culture and type of emotion expressed (Mesquita, 2022, p. 147).

Cultural interactions with emotion are particularly important for L2 speakers living away from their home cultures. They must navigate emotions that are perhaps different or expressed differently, and then adapt to the cultural norms associated with these. Pavlenko (2014) described the burden on language learners:

> To move beyond initial and often faulty assumptions and to understand the emotional world of their host community, L2 learners … have to puzzle out unfamiliar behaviors, to identify what triggers which "emotions" and when, to learn how particular "emotions" may be managed and to discover what cues to pay attention to and how to interpret verbal and non-verbal "emotion displays." (p. 247)

Adapting to these cultural differences may take a lifetime or may never even be reached at all. Mesquita (2022) further described the behavioral interplay between individuals on a social and cultural level as a type of dance:

> Like partners in a dance, your emotions and those of others complement and steer each other to form the interaction. And shared cultural knowledge, in the form of language and practices, orchestrates the ways in which different individuals do emotions together. It is like dancing the tango at the rhythm of the music, together with a partner who knows their dance steps, as you know yours. The dance emerges from everyone knowing their moves, and from the moves being in sync with the music. Doing your emotions in a way that fits with the relationships in your culture, and with your position in those relationships, is akin to having the right dance steps. (p.164)

When L2 learners are out of step and show differences in the encoding and decoding of emotion through nonverbal behavior, there is the potential for breakdowns in intercultural communication. These

breakdowns may then straddle the line between emotion and its physical manifestation through nonverbal behavior.

Affect and second language proficiency

Cognitive mechanisms driving SLA and impacting second language assessment have been and continue to be a focus for applied linguists. Likewise, especially in the language testing literature, features of speech and writing that characterize particular proficiency levels have played a major role in the validation and refinement of instruments that are better able to track the development of language proficiency. Nonetheless, affect has been hypothesized as having a facilitating or limiting effect on test takers' responses, determining "not only whether they even attempt to use language in a given situation, but also how flexible they are in adapting their language use to variations in the setting" (Bachman & Palmer, 1996, p. 65). In recent decades, affective traits making up learner individual differences, such as anxiety, motivation, and learner attitudes, have received substantial attention in the literature, and other complex attributes such as willingness to communicate and grit have been hypothesized to further deepen our knowledge of the mechanisms of language learning. Perceptions of affect, as well as the emotions learners report to feel, have received somewhat less attention in the literature, though this is growing (Dewaele & Li, 2020; Prior, 2019). It is important to uncover their effects because shifting affective stances may impact test performances, which may then translate to test score variance. Also, although nonverbal behavior is often decoded through an affective lens, behavior and affect may have varied interpretations by different interlocutor-listeners.

The most frequently studied affective phenomenon in applied linguistics is most likely anxiety (Gkonou et al., 2017). Anxiety—"the subjective feeling of tension, apprehension, nervousness, and worry associated with an arousal of the autonomic nervous system" (Horwitz et al., 1986, p. ii)—has been studied as a trait-based, more permanent disposition (e.g., test anxiety; Scovel, 1978). It has also been regarded as a state, understood as an affective response to a particular stimulus (Spielberger, 1983). Higher levels of anxiety have been found to hinder the processing of language input, thus having deleterious effects on language output in speakers (Gardner & MacIntyre, 1993; MacIntyre & Gardner, 1994) and possibly

delaying language development (Dewaele, 2010). In speaking situations, it has been found to impact a range of output variables, serving as a strong predictor for subjectively scored temporal measures of fluency (Pérez Castillejo, 2019). It can even cause changes in how listeners react to speakers verbally and nonverbally (Gregersen et al., 2014). Nagle et al. (2022) argued that anxiety may be distracting for interlocutor-listeners, thus interfering with processes of comprehension, and resulting in decreased comprehensibility. It is no surprise, then, that it has been found to share a negative relationship with L2 proficiency, competence, and achievement (Botes et al., 2020; Clément et al., 1980; Clément & Kruidenier, 1985; Dewaele & Alfawzan, 2018; Dewaele & Li, 2022; Dewaele & MacIntyre, 2014; Dewaele et al., 2019; Jiang & Dewaele, 2019; Jin et al., 2017; Li et al., 2020; MacIntyre et al., 1997; Teimouri et al., 2019).

Closely related to anxiety is the notion of confidence. Although not direct opposites, anxiety relates to discomfort and tension during L2 use, and confidence, especially L2 *self*-confidence, "corresponds to the overall belief in being able to communicate in the L2 in an adaptive and efficient manner" (MacIntyre et al., 1998). Confidence has been found to explain large amounts of variance in various educational, occupational, and other performative outcomes (Ahammer et al., 2019; Cobb-Clark, 2015; Heckman et al., 2006; Judge & Hurst, 2007; Stankov et al., 2012). Stankov et al. (2012) argued that confidence can be an affective state, trait, or disposition, lying somewhere between a measure of cognitive ability and a dimension of personality. Self-confidence formed a core predictor of communicative competence in Clément's (1980) and Clément and Kruidenier's (1983, 1985) socio-motivational models of L2 proficiency. In these models, positive and frequent contact with speakers of the L2 was understood to build speakers' self-confidence, leading to greater communicative competence, acculturation with the target L2 group, and adaptability (Clément et al., 1994; Noels & Clément, 1996; Noels et al., 1996). Confidence, then, was considered a core factor of motivation, and the development of both led to proficiency gains (Labrie & Clément, 1986), leading Clément (1986) to claim that self-confidence was "the best predictor" of proficiency (p. 286). Confidence can also lead to biases in the self-assessment of proficiency, with low confident speakers underestimating and highly confident speakers overestimating their abilities (MacIntyre et al., 1997). The relationship between confidence and L2 proficiency, however, may be cyclical, with

bidirectional feedback between the two as learners' language skills develop (Edwards & Roger, 2015). That is to say, confidence may help speakers perform in a second language (e.g., Doqaruni, 2015), and it may help them be perceived as more capable, while at the same time enhanced language development likewise may very well boost these learners' confidence.

Positive emotions have also received attention in the literature, drawing inspiration from interest in positive psychology (Fredrickson, 2001, 2003; MacIntyre et al., 2019; Seligman, 2011). Positive psychology considers the roles of positive emotions and affective stances in people's lives, and how these feelings may impact subsequent performances, achievement, and overall development. Positive feelings, such as happiness, warmth, enjoyment, and well-being play a critical role in language learning (Oxford, 2016). As opposed to the restrictive, debilitating effects of anxiety, positive emotions can enhance language learning by "broadening a person's perspective and opening the individual to absorb the language" (MacIntyre & Gregersen, 2012, p.193). Enjoyment has received the bulk of the focus amongst positive emotions in the L2 literature, though positive psychology has close connections to other extensively studied phenomena such as motivation (MacIntyre et al., 2019). Dewaele and MacIntyre (2014) argued that enjoyment and anxiety are not direct opposites; each has distinct functions regarding language learning and acquisition and different characteristics. In a handful of studies, both anxiety and enjoyment were analyzed in relation to L2 competence and achievement, with results indicating a positive relationship between achievement and enjoyment (Botes et al., 2020; Dewaele & Alfawzan, 2018; Dewaele & Li, 2022; Dewaele & MacIntyre, 2014; Dewaele et al., 2019; Jiang & Dewaele, 2019; Li et al., 2020). Li et al. (2020) was careful to note that any relationships between enjoyment and achievement/competence may be bidirectional. That is, enjoyment may drive language development similar to confidence, and language development may simultaneously drive enjoyment. It is also notable that studies analyzing the relationships between anxiety/enjoyment and achievement/proficiency were correlational in design, making the direction of causality less clear.

Engagement is another positive affective stance that helps to regulate interpersonal communication. Engagement has been broadly defined as an individual's level of interest and evidence of participation in

an event (Philp & Duchesne, 2016), and it may be composed of cognitive, social, and emotional elements (Baralt et al., 2016). Engagement may be strengthened by communicating on familiar rather than unfamiliar topics (Qiu & Lo, 2017), especially ones related to a speaker's life or personal experiences (Lambert et al., 2017). It may also be higher when learners speak to higher proficiency interlocutors (Dao & McDonough, 2018), and in face-to-face, human-mediated tasks (rather than computer-mediated) (Baralt et al., 2016). Engagement has also been classified as an integral strategy in interactional communication (Zhu et al., 2019), and has been found to underscore many interactional phenomena such as agreeing, disagreeing, and managing turn-taking (Goturk & Chukharev-Hudilainen, 2023). In language testing contexts, engagement likely drives participation in communication, which intuitively would lead to enhanced performance. The need to show participation despite underlying communication problems has been documented, as some learners have been shown to avoid clarification sequences when they fail to understand their interlocutors by using minimal responses (e.g., simple backchannelling like nodding, saying yes) (Ducasse, 2010; Lam, 2015, 2021; Luk, 2010). These face-saving responses likely serve to show the test takers as engaged even though they are in reality feigning understanding.

Affect and interpersonal perceptions

Until this point, affective responses have been presented as phenomena that arise either within or between people and dependent on the cultural background of the individual(s). Affect has also been shown to correlate with language achievement, proficiency, or competence, where negative affect (e.g., anxiety, disengagement) tends to align with lower proficiency, and positive affect (e.g., confidence, enjoyment, engagement) may be more characteristic of higher proficient speakers. These correlations may reveal broad trends amongst learners, but they are not likely to be causal. A highly proficient speaker can be anxious during an unfamiliar task, and a low proficient speaker may enjoy speaking a language despite facing communication difficulties. A question arises then about the impact of these affective displays on other individuals. As emphasized throughout this paper, communication rarely occurs in a vacuum. Listener-interlocutors ultimately perceive the affective responses of others, and these perceptions may furthermore play a role in how the speaker's language proficiency is perceived.

As discussed in the previous section on nonverbal mimicry, people may produce similar or opposing behaviors when seeing the behaviors of their interactants. Research in psychology has also discussed the ways that the affective displays of one person may impact another. People may converge in their emotional responses, such as when excitement spreads amongst a team, or when someone feels empathy for another's pain. People may also diverge in their responses, such as with feelings of schadenfreude (gloating at someone else's disappointment). Likewise, one person's emotions may provoke a non-emotional response, such as an observation of one's behavior. In cases of convergence or divergence, people's orientations towards events may be the same, but not necessarily so; likewise, similar orientations may provoke different responses, such as opposing teams watching a sports game.

The emotional influence of one person on another may be due to an *emotional contagion* (Elfenbein, 2014; Hatfield et al., 1994), contracted when people converge on an undirected emotion (an emotional display without a clear source) immediately and automatically. In other words, the mere presence of a feeling, such as fear, spreads quickly and without appraisal of any particular stimulus (Parkinson, 2011; Parkinson & Simons, 2009). In Hatfield et al.'s (1994) original conceptualization of *primitive* emotional contagions, seeing the speaker/source's emotional expression through facial and bodily behavior leads to nonverbal mimicry and finally emotional matching in the listener/viewer. Related to these behavioral imitations, emotional mimicry (Hess & Fischer, 2013) involves people producing behaviors that correspond to convergent emotions. Emotional mimicry differs from a contagion in the viewer's interpretation of the affective meaning at play, as mimicry involves a degree of awareness and contagions happen unconsciously and without appraisal. Mimicry is more common when the two or more individuals have coordinated perspectives on the context of the situation (Bavelas et al., 1986; Bourgeois & Hess, 2008). Importantly, Parkinson (2018) noted:

> Mimicry is not an instinctive and automatic process that guarantees facial matching under all circumstances, but instead relates to communicative goals that already imply an emotional orientation to the other person. Emotional mimicry depends on emotional meaning rather than producing it. (p. 166)

Regardless of the underlying processes involved in contagions, mimicry, or social appraisal (Lazarus, 1991), what is clear is that people see others' affective reactions, and these may sometimes produce a convergent response in the viewer, automatically or after inferring the underlying causes.

Research has also considered whether affective contagions may alter not only the way a receiver may feel, but also perform. In a study of the impact of emotion on test outcomes, Lochner (2016) hypothesized that positive psychological states would result in enhanced performance on a reasoning test in comparison with negatively valanced emotions, and also that anger would result in better performance than sadness. She conducted pre- and post-intervention reasoning tests with 429 participants. In the intervention, she induced five general emotions from the participants: joy, sadness, anger, contentment, and a neutral state. She measured the emotional states of the participants after the intervention and found that emotions had successfully been transferred. However, contrary to findings from the literature, she found no effects of emotion on reasoning test performance. She concluded that reasoning may be less susceptible to emotional interference in a laboratory context, but it may be more noticeable in other types of performance.

There is evidence of interpersonal affective impact in the L2 literature, often occurring dynamically as conversations unfold. Negative feelings, such as anxiety, may be provoked by a speaker's interlocutor (Hashemi, 2011). These feelings may fluctuate when speakers encounter interlocutors with differing social status and familiarity (Shirvan & Talebzadeh, 2017). Affective responses may also vary depending on how individuals perceive their partner's level of engagement, such as with their choices of forms and frequency in backchanneling (Cutrone, 2005; Heinz, 2003; Lindberg et al., 2022). When a listener hears "mhmm" from an interlocutor, they may interpret this backchannel as an acknowledgement of what they said, as impatience, or even as an interruption (Cutrone, 2005). These various interpretations can be a source of miscommunication, leading to growing anxiety in the listener (Li, 2006). Low engagement in an interlocutor, such as appearing disengaged or uninterested, may also provoke autonomic responses of anxiety (Lindberg et al., 2022). Anxiety may surface when individuals talk to a more proficient speaker of a language (Sevinç, 2018) due to linguistic insecurity (Heng et al., 2012). In L2 testing settings, the perceived affect of examiners may sometimes alter the performance (and underlying psychological

response) of test takers (Briegel-Jones, 2014; Plough & Bogart, 2008). Viewing nonverbal behavior may drive emotional contagions or mimicry in the transfer of affect (Blairy et al., 1999), but not necessarily so. The impact of viewing affective displays may be attenuated by perceptive ability. Lindberg et al. (2022), for example, found no evidence of an emotional contagion of anxiety in L2 dyads during a speaking task. These researchers used a correlational design to compare physiological measurements of anxiety (using galvanic skin response) with perceived ratings of anxiety of their interlocutors. They found no relationship between the two. They speculated that the lack of contagion may have been due to a lack of being attuned to their partners' feelings. Thus, sensitivity towards others may be an important variable that drives these effects. While this study was conducted in person, it is unknown whether asynchronous online stimuli could also provoke congruent affective responses.

Affective responses can occasionally then be a driving force of emotional changes in others. There is also evidence that affect not only impacts how others feel, but how they perceive the person that displays the affect in other ways. That is to say, if a test taker is perceived as being happy, this happiness may then color the examiner's perceptions of *other* characteristics and judgements about the speaker, such as their competence. In a study of 200 trait and state notions from 12 dimensions related to organizational psychology, Wojciszke et al. (1998) found that holistic judgements of individuals were predicted almost entirely by morality (defined as an affective, emotional character, e.g., warmth, friendliness, positive affect) and competence (e.g., skillfulness, ability), with morality comprising the bulk of the variance. Drawing from this study, empirical work on social cognition (e.g., Fiske et al., 2007), and their own empirical research in social psychology, Cuddy et al. (2007, 2008) created the Behaviors of Intergroup Affect and Stereotypes (BIAS) map of affective impact whereby warmth and competence formed two key dimensions, shown in Table 2.2. Here, high competence and high warmth in a speaker may elicit feelings of admiration from a listener. One can imagine a confident, competent, friendly L2 speaker being received with admiration due to their personability. On the other hand, a competent yet cold individual may provoke feelings of envy, such as the feeling when one may feel another does not deserve certain success due to their lack of friendliness. When someone is seen as warm yet less competent, this may result in feelings of

pity for that individual, such as a happy student that fails an important exam. Finally, cold, less skillful individuals may be seen as a burden and treated with contempt. The use of nonverbal behavior that conveys desirable dimensions of warmth and competence may then benefit less advantaged social groups, perhaps even L2 users, by overcoming judgmental biases (Cuddy et al., 2011). Feigning competence through power posing—intentionally putting on a performative confident stance—can even result in desirable outcomes, such as being perceived as stronger during job interviews (e.g., Cuddy et al., 2015).

Table 2.2
Affective Impact of Warmth and Competence (Adapted From Cuddy et al., 2007, 2008)

	High warmth	Low warmth
High competence	Admiration	Envy
Low competence	Pity	Contempt

Far less is known about the impact of affect and emotion on listener-raters in the L2 context. There is anecdotal evidence of raters noticing and reporting various dimensions of affect in the testing literature, especially in the context of interaction (Ducasse & Brown, 2009; May, 2009; Nakatsuhara et al., 2021a; Orr, 2002; Sato & McNamara, 2019; Thompson, 2016). Ducasse and Brown (2009) reported raters' observations about candidates' interactional competence. They found that attention and engagement were important affective stances during interactive listening, noting their critical role when displaying comprehension or repairing breakdown with their speaking partners. Nakatsuhara et al. (2021b) speculated that the significantly higher number of clarification sequences in their contrast of video conferencing tests with face-to-face tests could have been due to a greater need to signal engagement and solve communication breakdowns in this format, as gestures and voice inflection may have been less salient online. The raters in May (2009) also appeared especially attuned to affect in the context of asymmetrical interactions, noting that engagement, attention, and confidence were perceived via body language to convey assertiveness. Raters drew on assertiveness to determine the effectiveness of interactions, which May (2009) cautioned "may be of concern, in that these characteristics could be seen as aspects of culture, and L1 usage" (p. 417).

In a subsequent study, May (2011) found evidence of a broader number of patterns indicating interactional competence, a number relating to affect. Raters noted, for example, that confidence, assertiveness, engagement, collaborativeness, and showing a desire to communicate were all positive characteristics of interaction. Theorists have also stressed that affective stances may be important in communication, such as confidence and empathy (Morreale et al., 2013), adaptability to cope with challenging situations (Harding, 2014), and patience, tolerance, and humility to negotiate for meaning in intercultural situations (Canagarajah, 2006).

The perception of affect, generally by seeing nonverbal behavior, can impact judgements of L2 proficiency, though empirical research in this area is quite scarce. Sato and McNamara (2019), as discussed earlier, found evidence for an orientation to affect amongst their raters. They found that composure and attitude made up 5.7% and 6.3% of the comments about the CET-SET and Cambridge exams, respectively. Among the features discussed, confidence was one of the most important affective displays that related to raters' perceptions of proficiency. Confidence was sometimes viewed through the use of mutual gaze and the absence of anxiety-related behaviors (e.g., self-adaptors, averted gaze, low expressiveness), and related to a perception of stronger communicative effectiveness. Socially oriented affect such as engagement was also a common observation that occurred during moments of collaborativeness, also factoring into judgements of confidence. Anxiety was perceived negatively, even in cases where the speaker's message was comprehensible, leading to judgements of lower competence. Raters were also attuned to interactional features, finding their presence (such as active listening, responding, backchanneling) related to being attentive, interactive, and collaborative, and also leading to greater perceptions of competence.

However, few studies have attempted to measure the impact affect may have on ratings of L2 speech. Three studies to date have considered subjectively perceived affect (Nagle et al., 2022; Ockey, 2009) and objectively measured affect (Chong & Aryadoust, 2023) and their impact on rated outcomes. Nagle et al. (2022) set out to measure the extent that anxiety and collaborativeness (which the authors noted was a proxy term for perceived social engagement) influenced scores of L2 comprehensibility. The authors analyzed data from twenty dyads taking part in 17-minute interactions. In each interaction, the authors made

repeated measurements of their own and their partners' perceived anxiety and collaborativeness, while also rating the comprehensibility of their partner. Both anxiety and collaborativeness correlated at roughly .50, and each predicted comprehensibility, explaining 59–60% of the variance in the ratings depending on the task. The authors noted that collaborativeness may have comprised various orientations towards the task, speech content, and interactional competence. Anxiety, the authors posited, was likely perceived through various behaviors reported in the literature, such as a lack of expressiveness, gaze aversion, and the use of self-adaptors (e.g., Gregersen, 2005; Lindberg et al., 2021). They noted that these visual cues "may have made processing the L2 speaker's message more effortful for the interlocutor, leading to lower comprehensibility ratings" (p. 12). Whether perceived anxiety and engagement can have an impact on other ratings of language, such as fluency, grammar, and vocabulary, is unknown.

Ockey (2009) investigated the role of assertiveness on performances in group speaking tests. Although assertiveness may be considered a personality trait, its discrete appearance in tests may also be interpreted as a type of interpersonal affect. In his study, the test taker's level of assertiveness impacted the subsequent scores the test taker received. Higher assertiveness related to higher scores, but only when assertive test takers were paired with less assertive test takers. When paired with other assertive test takers, on the other hand, the effect was reversed. Ockey's study highlighted the importance of viewing affect or personality through the lens of context, as merely displaying one affective stance was not enough to change outcomes unilaterally. His study also highlighted the co-constructed nature of interactional performance in group test settings. Nonetheless, the impact of affective personality traits on score outcomes has been widely contested with little consensus to date (Berry, 2007; Davies, 2009; Nakatsuhara, 2011, O'Sullivan, 2004).

Chong and Aryadoust (2023) investigated emotional transfer and impact on L2 performance in an online speaking test. They exposed sixty undergraduate students in Asia to L2 English speaking performances in audio-only and audiovisual modes evoking either sadness or happiness as determined using sentiment analysis. There were thus four conditions, audio-happy, audio-sad, video-happy, and video-sad. Afterwards, raters asked them comprehension questions about the stimuli. The comprehension questions

were rated by four trained raters using the TOEFL iBT integrated speaking rubrics, which were (and are) available online (Educational Testing Services, n.d.). Chong and Aryadoust then analyzed videos of the test takers reactions to the videos using FaceReader 8.0 (FaceReader, n.d.), an automated, machine learning software that measures emotional facial behavior. They analyzed seven basic emotions: happiness, sadness, anger, surprise, fear, disgust, and a neutral state. They found that the happy videos indeed induced happy responses in participants' facial behaviors, and sadness induced negative valence emotions such as disgust. Videos produced a higher intensity of emotions than audio-only stimuli. However, participants' scores did not change as a result of modality or emotional state in the stimuli on any of the rating categories. There were, however, low correlations between the participants' actual measured emotional reactions and their performance outcomes, though the authors did not interpret these possibly due to the complexity of the findings, conflicting correlations, or low variance explained. The authors concluded that there was little to no effect of the test taker's visible emotions on their rated performance.

Measurement of affect and nonverbal behavior

There are many considerations in the use of nonverbal behavior and affect as either independent and dependent variables in research, and it is beyond the scope of this literature review to list them all. Reviews such as Gray and Ambady (2006) and Harrigan (2013) provide important methodological considerations for the use of different stimuli (images, thin slices, videos, interactions) to document behaviors. Generally, however, there are three methods for retrieving and coding nonverbal behavior: observational methods, physiological methods, and software methods. There are examples of each method in the L2 literature, and authors have used these somewhat differently.

Observational methods

The oldest form of studying behavior involves observational methods, in which a researcher annotates what they see in some written or numerical form. One common annotation form is in the form of discourse or conversation analysis, where speech is transcribed and nonverbal behaviors are indicated at points in the transcript where the behavior occurs. These annotations generally only indicate the onset of a behavior and not the duration and may be standardized or unstandardized. One of the most well-known

forms of gesture annotation is the McNeillian system (McNeill, 1992). This system allows for the description of the occurrence of gestural onsets and releases ([ ]), holds (…), and other phenomena such as filled pauses and silences. Gestural types are described in line. An example of this system is provided in Figure 2.4 from McNeill (1992, p. 95). The text example is of a monologue narrative in which a participant is recounting a story. Various gestures are identified, such as a beat gesture between lines 1 and 2, and the location of the gesture occurs within the bracketed word [ok]. Other gestures are described more fully. This type of system has been used extensively by gesture researchers such as Marianne Gullberg, and a similar system was used by Jenkins and Parra (2003) and Gan and Davison (2011). While this system is powerful for indicating the location of co-speech gestures, its use with other behaviors appears to be quite limited, as their timing and duration are sometimes less clear.

Figure 2.4
McNeillian System of Coding (McNeill, 1992, p. 95)

1. [ok] / (.4) and it was <was> (.3) a cartoon called Canary Row ǂ (.8)

Beat

2. <and> (.3) it was Sylvester and Tweety Bird ǂ (.3)
3. <and uh> (1.0) / (.1) it starts out with Sylvester sitting in
4. the Birdwatcher's Society
5. whi[ch is way] up in a building several floors up

O-VPT iconic: hands rise up at center to show height.

6. and he looks out with binoculars
7. and he sees Tweety Bird across the street
8. on a [window sill] ǂ (.4) in his little bird cage ǂ (.4)

O-VPT iconic: hand flattens and pats down to show window sill.

9. and Tweety Bird has binoculars
10. and he looks back
11. and ["I tawt I taw a putty cat"]

C-VPT iconic: head shakes back and forth. (voice changes)

Other systems annotate behavior in a separate line from speech in an attempt to align their temporal cooccurrence. As an example, Neu (1990, p. 134) used the Foster system of discourse annotation (Foster, 1980, as cited in Neu, 1990) to detail eye gaze behavior in Figure 2.5. In this example, "T" indicates that gaze is in the direction of the conversational partner, "c" is the ceiling, "r" is right, "l" is left, and "d" is in the general direction of the interlocutor, and "desk" is looking at the desk. In the same line of annotation,

"b", "r", and "up" indicate various movements of the head. The example effectively shows how gaze shifts and head turns unfold during the spoken sequence, giving the reader an understanding of how the two behaviors may align with speech and help manage interaction. There are many disadvantages of using such as system, however. For one, gaze and head movements occupy the same analytical line and may thus be confounded. It is unclear how coincides of various behavior would be annotated. Speech articulation rate is also not given, and as such segments of the discourse do not represent scaled moments in real time. While the duration of behaviors across words can be understood, it is unclear how much time these durations actually occupied. Other systems, such as the Birdwhistle system (Birdwhistle, 1970) and Ochs system (Schieffelin & Ochs, 1979), are similar but take different approaches to the representation of time and other behaviors.

Figure 2.5
Discourse Analysis From Neu (1990, p. 134)

YEAH, UH, I THINK (UH) I FORGET HIS NAME UH FRED YEAH,
. c . . . . . l . . . . . T . . . . . .
FRIED (IT UH) HE TOLD ME ABOUT HOW HOW WE UH USE UH TAPE UH
.. d . . . . . T. . . . . r . . . b.
AFTER THAT (UH) CLASS ((clears throat)) (I THINK
. . . . . T . . . . . desk . . . . . .
UH CHOOSE) I HIS NAME BUT UH HE TOLD ME UH GIVE US HOMEWORK.
. . . . . T. . . . . up . . . . . T. . . . . .

Similarly, studies of sequential interaction have used multimodal conversation analysis to document the cooccurrence of nonverbal behavior between interactants. In these cases, because of the complexity of the interaction, generally researchers choose a small set of behaviors to document in the unfolding sequence. For example, in Figure 2.6 (Seo & Koshik, 2010, p. 2226), a conversational exchange is documented between two individuals, after which a sequence of nonunderstanding occurs. The researcher here has documented the ensemble of behaviors (head, eye, eyebrow, and posture movement) at the point the behavior occurs with a description. This type of analysis is useful when a particular appearance of a behavior at a particular moment is salient, providing the researcher and reader with an emic perspective of micro-level features of interaction management. Nevertheless, information about the behaviors occurring before and after, as well as the duration of the ensemble, are not possible, and a more complete view of the

coincidence with other behaviors is lost.

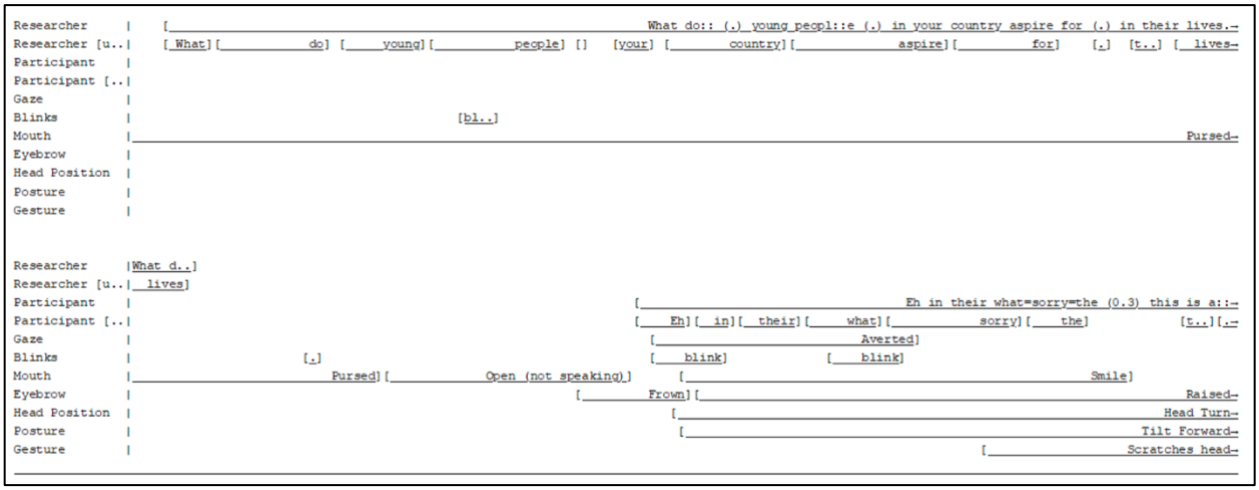
Figure 2.6
Multimodal Conversation Analysis From Seo & Koshik (2010, p. 2229)

```
SH:             so I would like to: .h visit: (.) I would
                like to visi:t, Kansas.
TL:             °m[ m°
SH:               [ en: Hahnnibal.
        ->      (0.5)/((TL tilts her head down to the left while
                        opening eyes wide and raising eyebrows;
                        then pokes her head and upper body forward;
                        all the while maintaining mutual eye gaze with SH))
```

An alternative type of multimodal analysis is possible using ELAN (EUDICO [European Distributed Corpora] Linguistic Annotator) software (Max Planck Institute, 2020). ELAN is open-source, free-of-charge audiovisual transcription software developed by the Max Planck Institute. ELAN has been widely used in the study of behavior in psychological and social research. It employs a tier-based annotation system that facilitates the coding of a wide range of spoken and visible phenomena aligned with the frame-by-frame unfolding of the video. It is thus possible to conduct a micro-level analysis of the timing and alignment of multiple behaviors with linguistic aspects of discourse, with various output styles, including multimodal conversation analysis. It is also possible to aggregate these timings to investigate the frequency and duration of each action or behavior over the course of analyzed segments. An example of such an annotation system is presented in Figure 2.6 (Burton, 2021a, p. 43). In this example, seven tiers of behavior have been annotated as well as four tiers of speech (two per interactant; one tier for conversation analysis, one tier for individual words). The example demonstrates the appearance of a nonunderstanding sequence and the unfolding of a hold. Because the annotated behaviors align temporally with speech and are scaled, it is possible to observe behaviors happening throughout the sequence and their alignment with trouble sources. In this particular case, a test taker has illustrated their lack of understanding by deploying an ensemble of behaviors at the same time, followed by a self-adapting gesture. This type of annotation is much more informative and transparent, but it can be extremely impractical due to the amount of time necessary to produce full annotations.

Figure 2.7

ELAN Multimodal Annotation System From Burton (2021a, p. 43)



Non-discourse analytic studies often opt to extract nonverbal behavior using raw frequency counts. For example, Tsunemoto et al. (2022) used a bottom-up approach to determine behaviors that appeared in a set of speech samples from a larger corpus. The authors watched videos several times, noted behaviors salient to discourse sequences, and then counted the behavior totals for each video. The advantage of this type of analysis is that it can be relatively quick, and interrater agreement can be fairly high. Using frequency data, however, only provides one piece of data about the occurrence of behavior. For example, if an individual smiles only once, but holds that smile for a long period of time, this instance of behavior would likely have a larger impact than a single smile held only momentarily. Yet, in this type of analysis, these frequency counts would be equivalent. Likewise, important information is lost about the location of behaviors. In non-linguistic studies, counts of individual muscle movements have been used, especially when documenting their relationship with emotion. Ekman et al.'s (2002) Facial Affect Coding System (FACS) has been used to provide a more complete picture of facial anatomy, including intensity and timing. Using FACS requires significant training and has been used in a number of studies outside of applied linguistics (Ekman & Rosenberg, 2005).

Each of these systems requires reliability checks to ensure the integrity of the annotations. Some studies have two individuals annotate the entire dataset, and any disagreements are discussed and resolved. This type of dual coding was reported by Jenkins and Parra (2003) and Tsunemoto et al. (2023). For more

complex analyses, however, resources are often insufficient, as annotations can last weeks at a time. Duncan and Fiske (2015) recommended that 10% to 25% of datasets be double coded prior to coding the entire dataset, after which disagreements can be resolved, and then the rest of the dataset can be coded by one individual.

Finally, observations of affect may be gathered using scales. Scales can be a quick and intuitive tool for observers to make decisions about affective phenomena. Nagle et al. (2022), for example, had raters use simple scales of collaborativeness and anxiety to score their own and their partner's affect. These scales were set on a sliding bar which represented a total of 100 points (one point per millimeter), allowing the researchers to use a bounded continuous variable as an independent variable. Similar scales were used in Tsunemoto et al. (2022) and Kim et al. (2023) to measure comprehensibility, accentedness, and fluency. Other scales may be set up in a Likert format, with a limited number of ordinal points, such as semantic differentials (Osgood et al., 1957; Snider & Osgood, 1969).

Physiological methods

In some cases, physiological methods may be preferable as an objective (or quasi-objective) measure of behavior. For example, eye tracking technology (Godfroid, 2019) can be used to track the eye gaze patterns of individuals when interacting with language when reading or listening, but it can also be used to track interpersonal communication. For example McDonough et al. (2015) used eye tracking to measure gaze location and duration during recast episodes. Likewise, Batty (2021) measured areas of the face that L2 users attended to when watching and listening to speakers in video-based listening tests.

Quasi-objective measures may also be informative when studying affect. These measures provide electrical information from the body that has been linked to either muscle stimulation or autonomic arousal. For example, facial electromyography (EMG) detects electrical signals from muscle activation that may not be otherwise visible. It can be used to detect signals occurring with particular expressions of emotions, but it is not generally used to detect the facial movements without prior knowledge of affect (Cacioppo et al. 2000). It is, however, quite obtrusive, requiring equipment on the face or even below the skin. Other physiological measures, such as galvanic skin response, can be much less intrusive. Galvanic skin response

is thought to measure autonomic arousal, which is associated with affective responses (for example, anxiety), but it is a rather crude measure with no direct relationship to any particular emotion (Afifi & Denes, 2013). As an example, Lindberg et al. (2021) used galvanic skin response as a proxy for anxiety, which allowed them to correlate their findings with the occurrence of certain nonverbal behaviors. A benefit of this type of measure is it is relatively unobtrusive, but the validity of its inferences may be limited to arousal rather than specific affective output.

Software

The final system of measuring nonverbal behavior and affect is through computer vision and machine learning. Decades of work have resulted in greater and greater accuracy in systems that can measure these indices without the participant's awareness, thus potentially enhancing the ecological validity of research studies. With software-based measurements, participants are recorded (often sitting in front of a computer with a front-facing camera) conducting a study online. Participants need not wear any equipment, as the software is able to extrapolate muscle movements from video alone. Another important advantage of these systems is that they offer speed and objectivity in measurement. These systems work by using machine learning to identify specific points on the face corresponding to Ekman et al.'s (2002) Facial Action Coding System. The software then produces probability measures of the activation of individual facial muscles. Using models trained on classified banks of individuals exhibiting certain emotions, the software then can produce a probability measure of the type and strength of various measures. These models, however, sometimes operate somewhat opaquely and have limited published validity evidence of their internal functioning, and it is also sometimes unknown to what extent these may work with individuals from varying cultural backgrounds. Additionally, the accuracy of emotional classification in a comparison of various automated facial recognition systems in 2020 showed that although better than chance, these systems were not as accurate as judgements made by human observers (Dupré et al., 2020).

Chong and Aryadoust (2023) authored the only study to date, at the time of writing, that has used an automated facial recognition system in the study of L2 proficiency. They used the FaceReader emotion recognition system (FaceReader, n.d.) to extract indices of happiness and sadness which they then analyzed

in comparison to language proficiency subscores. FaceReader (FaceReader, n.d.) is able to produce indices of seven basic emotions and individual action units pertaining to facial muscle movements. Although FaceReader was purportedly more accurate than many other systems (in 2020; Dupré et al., 2020), it fails to recognize many emotions identified by human observers (Hirt et al., 2019). Additionally, FaceReader is unable to produce omnibus measures of overall expressiveness, attention, or a general measure of emotional valence. The literature to date does not point to a direct effect of specific emotions on proficiency outcomes but rather overall expressiveness and positivity (Jenkins & Parra, 2003), more complex affective measures (e.g., anxiety and engagement; Nagle et al., 2022), or specific behaviors (e.g., eyebrow movements and smiles; Kim et al., 2023; Tsunemoto et al., 2022).

An alternative system that addresses some of these limitations is iMotions Affectiva (iMotions, 2017). iMotions is a behavioral analysis application that uses a complex array of computer vision and machine learning algorithms to detect faces, automatically code facial expressions, and classify emotional states. iMotions is able to detect head orientation, facial landmarks, action units or expression metrics, seven emotional states (joy, anger, surprise, fear, contempt, sadness, and disgust), and three omnibus measures of valence, engagement, and attention (iMotions, 2017). Although Dupré et al. (2020) found iMotions Affectiva to be somewhat less accurate than FaceReader, the software has been found to be more accurate than physiological methods such as facial EMG (Kulke et al., 2020). Only one study in applied linguistics to date has used iMotions software to my knowledge, but it used the eye-tracking component with no facial behavior analysis (Suvorov, 2015).

Summary

The studies I have reviewed suggest that there is a complex and natural relationship between nonverbal and verbal communication. The two modes may offer complementary semantic information that strengthens speakers' intended messages. When engaged in decoding meaning, individuals are informed by these two physically separate yet conceptually intertwined lines of cognitive, affective, and social information. With judgements in performance tests, raters may use the nonverbal information available to them to complement their understanding of someone's L2 proficiency, thus making use of external

nonverbal criteria to decide whether a performance merits a higher or lower score. If a low proficiency speaker uses nonverbal behavior to express strong positive affect, or if a higher proficiency speaker does not display much affect at all, it may be the case that nonverbal information plays a role in decisions about language proficiency. In these cases the direction of the effect of nonverbal behavior may be positive in the first case and negative in the second case. It is also possible it may have no effect at all. The effect of nonverbal information on judgements of verbal ability may also not be uniform for all proficiency levels or all criteria included in a rating scale. Scant research informs these areas of inquiry.

Overall, research studies on the effects of nonverbal behavior on raters' scores are inconclusive. Simply comparing mean differences between rating samples with video and without video may mask underlying interactions between raters, proficiency levels, and criteria scores. For example, one rater in Nakatsuhara et al. (2021a) was found to exhibit severity on the video mode, while other raters showed more severity on the audio-only mode. If this type of differential severity were split evenly in a group of raters, any mean group differences would possibly be cancelled out. Nonetheless, evidence largely points to a benefit of the visual mode of rating on scores, but it is unknown whether the benefit, ostensibly due to nonverbal behavior, is consistent across proficiency levels. It is important to consider such interactions where possible.

Studies that took a fine-grained, discourse analytic perspective provided evidence that indeed raters appeared to behave in similar ways documented by Burgoon et al. (2016): When test takers' linguistic skills fall into borderline categories, or if nonverbal information conflicts with verbal, nonverbal communication may take precedence when raters make decisions about L2 ability. In particular, expressiveness and attentiveness (mutual gaze) were cited as important criteria that brought borderline scores up, while relative inexpressiveness and inattentiveness had the effect of bringing scores down. Gestures, posture, and paralinguistic features also contributed to raters' impressions. In all studies that asked raters to provide reports of their decision-making processes, nonverbal behavior was often mentioned through its functional output as an affective response. Raters were sensitive to behaviors that conveyed confidence, anxiety, engagement, attention, interactiveness, and positivity, especially when these expressed a desire to

communicate.

To date, the studies reviewed here have either taken a small scale, discourse-analytic approach or a larger scale, score-based approach. The gap here, then, is to be able to analyze scores with a larger bank of raters and a more diverse sample of test taker behavior, and to produce a rich amount of qualitative data that can triangulate findings between scores, behaviors, and rater comments. In the past, this type of analysis would have been impractical: Detailed behavioral analyses of even one minute of speech can take hours. Today, it is possible to leverage machine learning technology to extract behavioral indices that can be used in statistical models to better understand the moderating impact of nonverbal behavior and affect on proficiency scores.

CHAPTER 3: RESEARCH QUESTIONS

The literature shows that nonverbal behavior is a core aspect of communication that conveys semantic, cognitive, affective, and social-interactional information. Speaking test constructs, including those based on communicative competence, have been shown to tacitly acknowledge nonverbal behavior, especially in its social-interactional roles of conversation management and strategic planning, but these lack a full exploration of the functional output of nonverbal behavior. Nonetheless, variance has been found in testing contexts that is attributable to the visual realm, with links both to the behavioral forms and the informational output, often as affect. Missing from the literature to date is an exploration of score variance across various language proficiency outcomes with a dataset large enough to measure the strength of associations. Likewise, there are methodological gaps in the measurement of affect and behavior, with most studies measuring few variables or in a limited measurement style, such as frequency. This dissertation attempts to bridge those gaps by answering three key sets of research questions.

Proficiency scores and interpersonal affect

This question block focuses on interpersonal affect. Using a range of perceived affective variables, the aim is to describe trends and patterns in the variables that may explain language proficiency outcomes. These variables will be observed by the raters in the study. There is one question in this section:

RQ1: What is the relationship between interpersonal affect and language proficiency?

No hypotheses were generated prior to formulating this question, as this question was purely exploratory. However, the literature points strongly to confidence, anxiety, and engagement correlating with language proficiency, so it is anticipated that some aspects of language proficiency would correlate with these variables. Less is known about the relationship between positive affective variables and outcomes. Research on foreign language enjoyment has suggested a link to L2 achievement.

Proficiency scores and nonverbal behavior

The second block of questions relates to externally measured variables of nonverbal behavior and language proficiency. As described in the methods, these variables will be measured objectively using automated software. Base proficiency measures will be obtained by using official scores from a testing

agency. These questions are confirmatory in nature.

RQ2.1: Do externally measured indices of nonverbal behavior predict language proficiency scores?

RQ2.2: Do nonverbal behaviors impact outcomes differentially depending on the base proficiency levels of test takers?

After generating these research questions, I proposed three hypotheses relevant to this block of questions, which I pre-registered in the Open Science Framework (Burton, 2021b).

H2.1.1: Indices of attention and expressiveness will have significant but moderate correlations with language ability.

H2.1.2: Higher values of attention and expressiveness will result in significant positive regression coefficients of fixed effects, indicating an overall positive impact on impressions of second language proficiency across ability levels.

H2.2: Significant interaction coefficients of base language proficiency with attention and base proficiency with expressiveness will indicate that the effect of nonverbal behavior on rated outcomes depends on the base proficiency of the test taker.

Raters and nonverbal behavior

The third block of questions is a mixed-methods inquiry into rater behavior. The primary goal of this block is to triangulate findings from the quantitative analyses. These questions are largely exploratory.

RQ3.1: Which nonverbal behaviors are most salient and informative to raters when scoring?

RQ3.2: How do raters understand language proficiency in light of nonverbal behavior?

One confirmatory hypothesis was pre-registered for question 3.1 based on a close reading of the literature reviewed previously.

H3.1: Gaze aversion, eyebrow raises, smiling, head tilts, and inexpressiveness will be mentioned more times by raters as noted by higher relative frequencies of comments. Gesture and posture will be mentioned fewer times due to the online format of the speech stimuli

CHAPTER 4: METHOD

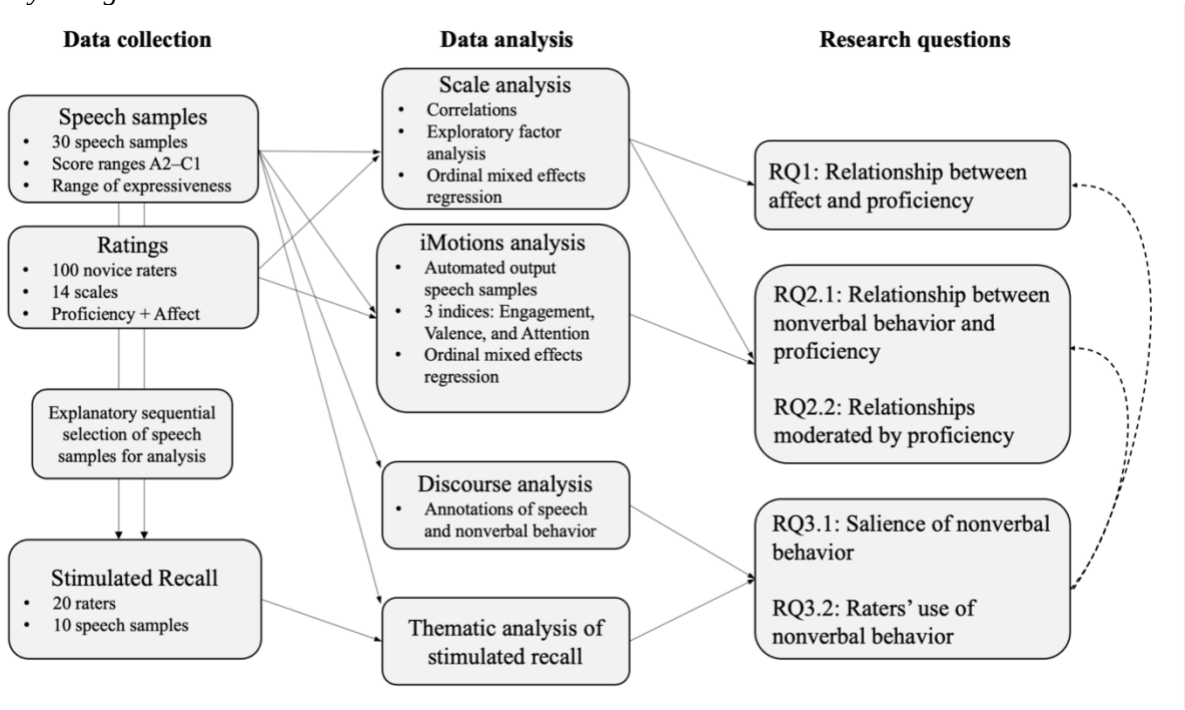
This dissertation is, at its heart, a study of speech perception, as the core questions surround how listeners make use of visible nonverbal behaviors and affect while thinking about speech. It is also a study of speech assessment, as the main method of the study is for participants to assign a score that represents a particular level of language proficiency. By analyzing scores, it may be possible to observe the direction and size of an effect of the visual realm on people's perceptions of language proficiency. Score analysis alone, however, will not be able to determine whether raters' use of nonverbal behavior when assigning scores is explicit (i.e., raters are aware of behavior during rating) or implicit (i.e., behavior impacts rating unconsciously). Uncovering this explicit/implicit distinction is only possible by probing raters' thought processes through rater reports. For this reason, this study is also one of rater cognition. In order to explore all of these elements, the study takes a mixed-methods approach.

I adopted an *explanatory sequential design* (Creswell & Plano Clark, 2017) as a mixed-methods framework for this study. In explanatory sequential designs, quantitative data is gathered first to identify cases which may be indicative of or exemplify the phenomenon under consideration in the research study. Following this, qualitative research methods are used to explain the nature of the quantitative results in much more depth and detail. Explanatory sequential designs are most appropriate "to explain the mechanisms through qualitative data that shed light on why the quantitative results occurred and how they might be explained" (Creswell & Plano Clark, 2017, p. 77).

The design of the study consisted of three main phases. Phase 1 took place in Fall 2021. It included two components. In the first, I piloted the online Qualtrics survey used by the raters, using data from my first qualifying review paper (Burton, 2023). In the second component, I selected the speech samples from those provided to me by IELTS. Phase 2 began in December 2021. This phase included the recruitment of rater participants and the collection of quantitative rating data of both proficiency and affect. This phase lasted until April 2022. During this phase, I selected speech sample stimuli that exhibited rating patterns suggesting a greater impact of nonverbal behavior on scores for further qualitative analysis. In this phase, I also collected automated visual output of the speech sample videos using iMotions (iMotions, n.d.). Phase

3 overlapped with Phase 2 in the last weeks of data collection. In this phase, 20 raters were invited to take part in stimulated recall sessions within 24 hours of completing the online rating study. These stimulated recalls targeted a subset of speech stample stimuli selected in Phase 2 of the study. Phase 3 took place during March and April of 2022. Together with two research assistants, I also transcribed and annotated nonverbal behavior in the speech samples to analyze together with the stimulated recall data.

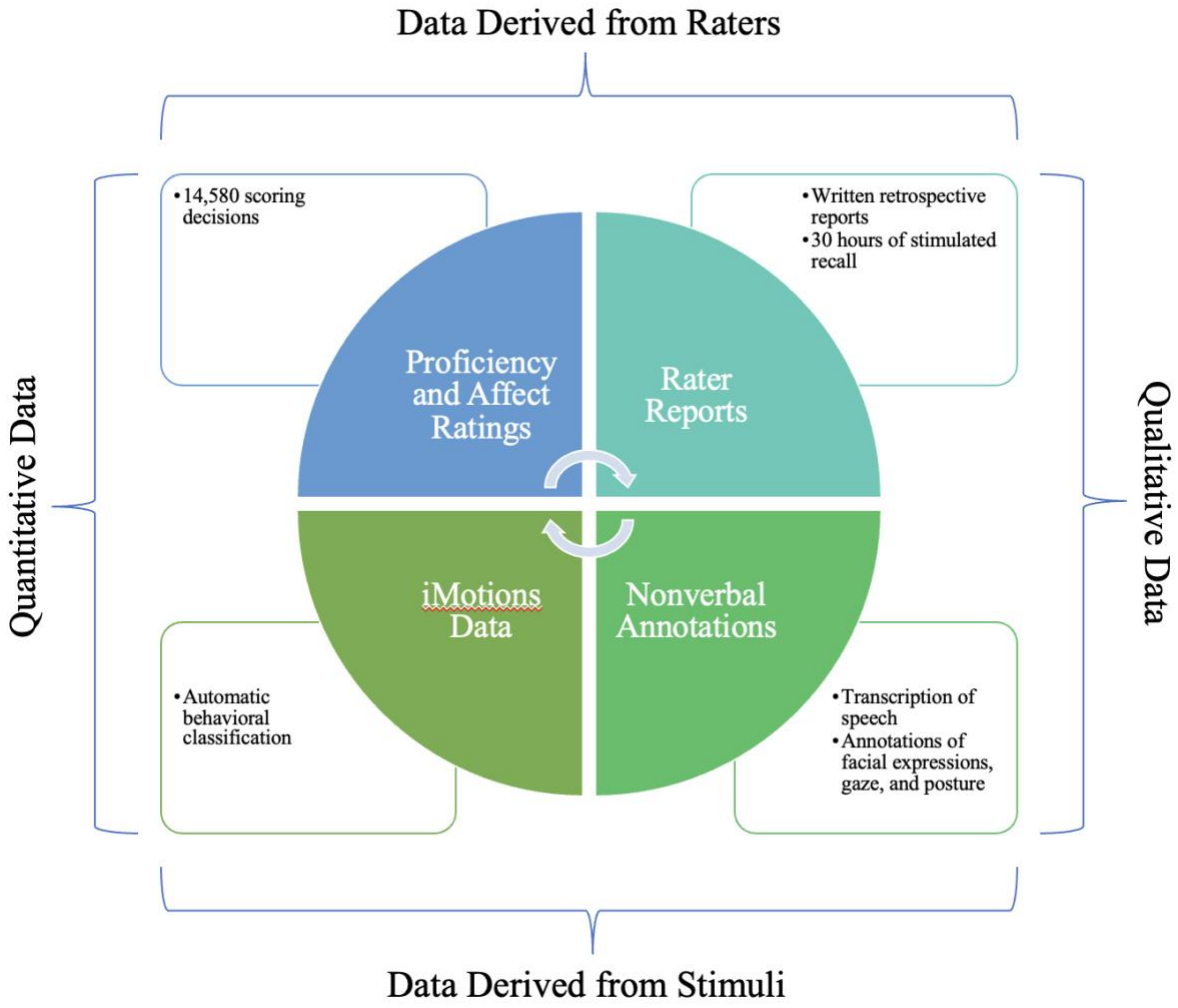
Figure 4.1
Study Design



The explanatory sequential design of the study is visible in Figure 4.1. The top half of the figure illustrates the quantitative aspects of the project, including the rating of speech samples with scales and iMotions analyses. The quantitative part of the study served to answer the first two blocks of research questions. The explanatory sequential mixed-methods aspect of the study is described in the bottom half. Here speech samples selected from the quantitative component formed the basis of the stimulated recall sessions, which were analyzed thematically, and together with samples of multimodal discourse served to answer the third block of research questions. This third block, however, was used to reflect back to the quantitative findings. The design of the study was thus triangulated. The dataset included sources of data from both raters and speech stimuli. the four main data sources Figure 4.2 further illustrates the four main

data sources and their relationship to the data analysis.

Figure 4.2
Data Sources for Dissertation



Participants

There were two groups of participants in this study. The first group of participants completed the rating study. A subset of this group comprised the smaller second group of stimulated recall participants. The two groups are described below.

Rating study

Participants for the rating study were recruited from a pool of undergraduate students at Michigan State University called SONA (https://psychology.msu.edu/undergraduates/sonaparticipation.html) and from the university's undergraduate student email list. Participants were allowed to participate as long as

they were traditional undergraduates (in the age range of approximately 18–22), born in the United States, and had grown up with English as their L1. These conditions were meant to ensure that participants had a similar background; that is, they stemmed from a shared cultural and linguistic context, as individuals from differing L1 or national backgrounds may vary in their interpretations of language proficiency (Barnwell, 1989; Kim, 2009; Marefat & Heydari, 2016; Shi, 2001; Wei & Llosa, 2015; Xi & Mollaun, 2011; Zhang & Elder, 2011) and may possess culturally or linguistically different nonverbal behavior and affect (Crivelli & Fridlund, 2018; Fridlund, 1994; Matsumoto & Hwang, 2016).

Recruiting young undergraduates also ensured that no individual had operational experience rating speaking tests or any familiarity with rating scales of language proficiency; that is to say, the rater participants were untrained, often referred to as *novice* or *naïve* raters, or *linguistic laypersons* (Sato & McNamara, 2019). The principal benefit of using untrained, novice raters in designs such as this study is that scores and other mediating variables extrapolate more strongly to the target language use domain:

> in the real-world context, the ultimate arbiters of L2 speakers' oral performance are typically not in fact trained language professionals, who have meta-level linguistic insight and are possibly concerned primarily with features of communication that are the focus of their own training as linguists or language teachers, but interlocutors with no specialist training (Sato & McNamara, 2019, p. 895).

The design of this study also included no rater training, as judgements were epistemically to be based on quick impressions based on the participants' general experience with interpersonal interactions in life. This in turn meant that scores would likely be less reliable and could exhibit major differences in rater severity (Attali, 2016; Barkaoui, 2010; Cumming, 1990; Lim, 2011; Shaw, 2002; Weigle, 1994, 1998). However, this variance was anticipated and desirable in this study, as it best reflects the way listeners in the target language use domain may process multimodal input.

To determine the sample size necessary to detect possible small effects in this study, I referred to previous literature on sample size requirements for mixed-effects designs (that is to say, designs where raters provide multiple observations per case). I also conducted a simulation study. Past literature suggested

that larger second level cases (in this case, rater participants) than first level cases (here, visual stimuli) would provide greater power (Hox et al., 2018), as long as the number of stimuli is large enough (Westfall et al., 2014). I ran the simulation analysis with 10, 20, 30, and 40 visual stimuli. I found that a stimuli size of 30 and a rating sample size of at least 80 would have a power of .95 to detect regression coefficients of .2 for the iMotions variables, which I considered the smallest meaningful effect size. This power analysis was similar to Westfall et al.'s (2014) finding of a standardized effect size *d* of .4 with a participant pool of near 100 and with 30 stimuli. In the power analysis I ran, smaller stimuli sizes required fewer raters to arrive at sufficient power (40 and 60, respectively), but reducing the stimuli size was not desirable for this study as there would be less variation in proficiency levels and nonverbal behavior. Using 40 speech stimuli would have required 120 raters to reach the same power, which would have furthermore increased the cost of the study. For these reasons, I set the desired stimuli size at 30 and the rater sample size at 100, as I felt this struck the right balance between desirable variance and practicality. It also allowed enough flexibility to drop problematic cases or outliers without losing excessive statistical power.

As referred to above, in order to recruit participants, I used both the SONA system and e-mail invitation blasts to all domestic undergraduate students on campus In total, 2,340 individuals signed up to take part in the study. I invited individuals that fit the requirements of the study randomly from this pool until I reached the targeted number of participants. In total, 281 individuals received invitations and 100 successfully completed the study. All participants completed a consent form and signed non-disclosure agreements in Appendix A. One individual experienced technical problems while completing the study, which resulted in incomplete data, so I removed this case from the dataset. The final dataset contained complete observations from 99 raters. The mean age of the participant raters was 20.92 years (*SD* = 1.48). Gender was relatively balanced, with 41% reporting identifying as male, 53% as female, and 6% as other (participants were not asked to specify further). All participants were from the USA and spoke English as their L1, and 38% reported speaking a second language, though participants were not asked to identify their level of proficiency or familiarity with their L2. School year was fairly balanced as well, with each year (freshman, sophomore, junior, senior) representing approximately 25% of the dataset. Finally, participants

reported studying in 37-degree programs, with three individuals reporting not having decided on their program at the time. The full demographic reporting is in Table 4.1.

Table 4.1
Demographics of Rater Participants

Category	*n*	Category	*n*
Gender		Degree	
Male	41	Accounting	1
Female	52	Biochemistry	1
Other	6	Business	1
		Communications	4
Nationality		Computer Science	3
USA	99	Creative Advertising	1
		Criminal Justice	1
L1		Economics	1
English	99	Education	2
		Engineering	12
Speaks L2		English	1
Yes	38	Environmental Studies	1
No	61	Finance	3
		Fisheries and Wildlife	1
Year of School		Genomics and Molecular Genetics	1
Freshman	25	History	2
Sophomore	22	Human Biology	7
Junior	21	Human Capital and Society	1
Senior	29	Humanities	1
Other (Non-traditional year)	2	Interdisciplinary	1
		Interior Design	1
		Japanese	1
		Kinesiology	4
		Linguistics	3
		Mathematics	1
		Neuroscience	9
		Nursing	5
		Packaging	1
		Physiology	2
		Psychology	7
		Social Work	4
		Spanish	2
		Supply Chain Management	4
		Theatre	2
		Zoology	4
		Unspecified	3

Stimulated recall

After 60 raters had completed the rating study, rating scores on language proficiency were used to identify stimuli to analyze further in the stimulated recall sessions. The process used to identify samples is described in a separate section below. I invited individuals that had not taken part in the study yet but had indicated being willing to participate in the stimulated recall sessions when signing up. I randomly invited participants until I reached a target sample size of 20. I chose 20 as this would represent a rather robust number of stimulated recall sessions at 20% of the total number of participants. A total of 42 participants were invited to participate in the stimulated recall sessions, and 20 successfully completed the sessions. Stimulated recall participants signed an additional consent form allowing them to be audio recorded; the consent form is available in Appendix A. The demographic breakdown was similar to the figures from the larger rater group. Demographics for the stimulated recall sessions are in Table 4.2.

Table 4.2
Demographics of Stimulated Recall Participants

Category	*n*	Category	*n*
Gender		Degree	
Male	8	Computer Science	2
Female	11	Education	1
Other	1	Engineering	2
		Finance	2
Nationality		Genomics and Molecular Genetics	1
USA	20	History	1
		Human Biology	1
L1		Japanese	1
English	20	Neuroscience	3
		Nursing	1
Speaks L2		Psychology	1
Yes	7	Physiology	1
No	13	Spanish	1
		Zoology	1
Year of School		Unspecified	1
Freshman	7		
Sophomore	4		
Junior	3		
Senior	6		

Materials

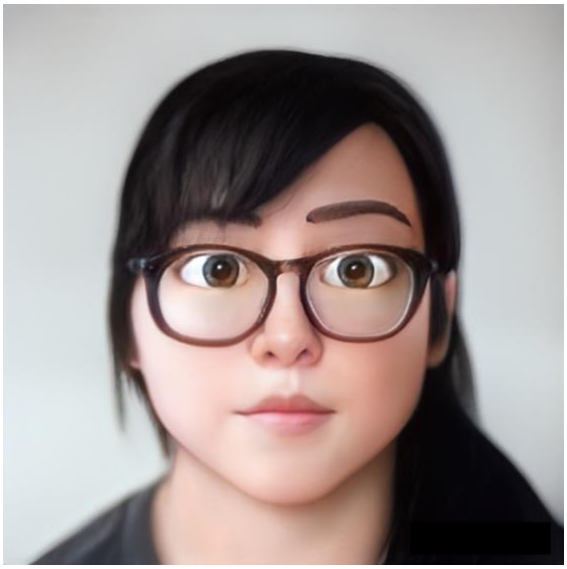
Materials used in this study included speech samples for rating, the rating scales, a survey of demographic information, a follow-up survey after the rating took place, the online Qualtrics platforms that hosted these materials, and a stimulated recall session for a subset of the raters using a subset of the speech samples. Each of these will be described below.

Speech samples: Rating study

The International English Language Testing System (IELTS; IELTS, n.d.) provided me with previously recorded, high-stakes, live speaking test samples for use as rating stimuli for this study. The videos were originally collected as part of the IELTS remote speaking research project (Nakatsuhara et al., 2017), and as such were recorded in operational settings as part of research into the feasibility of online speaking tests delivered through Zoom. IELTS is an international, high-stakes academic English proficiency test used primarily by universities and colleges for admissions purposes. IELTS includes four sections corresponding to the language abilities of reading, writing, speaking, and listening. The speaking test is composed of three parts. Part 1, which lasts 4–5 minutes, includes a basic introduction and personal questions. The format of Part 1 is that of an oral proficiency interview; the examiner asks scripted questions followed by the test taker's answers, without scaffolding or elaboration on the part of the examiner. This part primarily targets familiar, non-abstract topics. Part 2 is in the format of a monologue. Test takers are given a task card and are asked to speak on their own, without support, for up to two minutes on a topic. These topics generally elicit familiar information (e.g., *tell a story about your life*) with generalized, abstract conclusions (e.g., *how does this apply to society in general?*). Part 3, which comprises the final 4–5 minutes of the test, is a semi-scripted oral proficiency interview. It is distinguished from Part 1 by having the examiner produce unscripted follow-up questions and allowing the examiner to use a broader range of functions with the test taker; the examiner may scaffold, gloss, probe, repair, and follow up on answers the test taker has given. These topics are meant to be unfamiliar to the test taker, and the concepts gradually grow in abstraction. For illustrative purposes, example test questions can be seen at https://takeielts.britishcouncil.org/take-ielts/prepare/free-ielts-practice-tests/speaking.

IELTS provided me 46 video recordings of part 3 of the speaking test for use in this study. Overall speaking test scores, as well as subscores of the four analytical scoring categories (grammar, comprehensibility, fluency, and pronunciation) for each test taker were provided. The dataset that IELTS provided for this study included a range of total speaking scores (an average of the four subscores) from 2.5 to 6.5 (M = 4.97, SD = 1.07, approximately A2–C1 on the CEFR). All test takers were young Chinese adults in Shanghai. The dataset skewed female (71%) and featured nine different examiners (staff members who administered the speaking test to the test takers) and seven test topics (Education, Websites, Communication, Success, Events, Leisure, and Travel). Prior research found no impact of examiner, test topic, or interaction of both on test scores (Nakatsuhara et al., 2017). All tests were conducted in the same controlled setting on standardized laptop computers. The tests were recorded using Zoom's internal recording feature, which displayed both the examiner and the test taker side-by-side. Test takers faced the laptop directly, which had a camera embedded in the top bezel, thus simulating direct eye gaze. In order to protect the test taker's identity, I have cartoonized what the stimuli input looked like for the raters in Figure 4.3. Note that the raters did not see cartoons, but the actual video recorded sample.

Figure 4.3
A Video Frame of Sample 30, Cartoonized to Protect the Test Taker's Identity



As mentioned above, I had decided to use 30 speech samples with the raters based on the simulation study. Thus, I devised a method to select 30 of the 46 available samples. I first watched each sample to

document the quality of each video and the overall expressiveness of each test taker. I first removed videos with extended periods of silence or videos in which language production was too limited to be meaningful for rating language. I then sorted the videos into three proficiency groups of (a) 3.5 to 4.5, (b) 5 to 5.5, and (c) 6 and 6.5 using the overall rounded IELTS scores. The scores are rounded because this is how the test producer calculates the total score in operational settings. A critical aspect of this research project was to observe the impact of nonverbal behavior on raters' scores, and for this reason it was also necessary to include a range of expressiveness and behavior for the raters to observe. I classified each video crudely as less expressive (0), moderately expressive (1), and expressive (2) based on the presence of active facial behaviors such as smiling, frowning, nodding and mutual eye gaze as well as gesture. To compile the dataset, I selected stimuli from each scoring category with a balanced distribution of expressiveness, and an overall balance of score distributions. This resulted in eight stimuli with scores 3.5–4.5, 14 stimuli with scores of 5 to 5.5, and eight stimuli with scores of 6 to 6.5 (M = 5.2, SD = 0.88). Furthermore, 12 of the samples showed a high degree of expressiveness, 11 showed markedly low expressiveness, and seven fell somewhere in between. A table listing the 30 resulting stimuli and their features is presented in Table 4.3.

Table 4.3
Speech Sample Data

Sample	Gender	IELTS score					Examiner	Topic	Expressiveness	Time
		FC	V	G	P	Total*				
S01	M	3	4	4	4	3.5	8	Education	Low	02:23
S02	M	4	4	4	4	4	2	Success	Low	02:31
S03	M	4	4	4	5	4	4	Traveling	Low	02:19
S04	F	4	4	4	5	4	4	Events	High	01:43
S05	F	4	4	4	4	4	6	Success	High	02:09
S06	F	4	4	4	5	4	7	Success	High	02:18
S07	F	4	4	4	4	4	9	Success	Low	02:11
S08	M	5	4	4	5	4.5	4	Communication	High	01:53
S09	F	5	5	5	5	5	1	Education	Low	02:00
S10	F	5	5	5	5	5	1	Success	Low	02:31
S11	F	6	5	5	5	5	2	Success	Low	01:46
S12	F	5	5	5	5	5	2	Communication	High	01:49
S13	M	6	5	5	5	5	6	Cinema	Mid	02:13
S14	F	5	5	5	5	5	6	Traveling	Low	02:02
S15	F	5	6	6	6	5.5	1	Success	High	02:22
S16	M	5	6	6	6	5.5	1	Communication	Mid	02:19
S17	F	5	6	6	6	5.5	1	Traveling	High	01:51
S18	F	5	6	5	6	5.5	2	Events	Mid	02:18
S19	F	6	6	5	5	5.5	2	Events	High	02:05
S20	F	6	5	6	6	5.5	5	Success	Mid	01:57
S21	F	6	5	6	5	5.5	5	Events	Mid	02:23
S22	M	6	6	5	5	5.5	6	Education	High	02:13
S23	F	6	6	6	7	6	2	Cinema	Low	02:20
S24	F	7	6	6	6	6	6	Communication	Low	02:43
S25	F	6	6	6	6	6	6	Cinema	High	02:14
S26	F	7	6	6	6	6	6	Communication	High	02:13
S27	F	7	6	7	7	6.5	3	Traveling	High	02:14
S28	F	7	7	6	6	6.5	4	Cinema	Mid	01:53
S29	F	7	7	6	7	6.5	5	Events	Low	02:22
S30	F	7	6	6	7	6.5	9	Traveling	Mid	02:15

Note. FC = Fluency and coherence. V = Lexical resource. G = Grammatical range and accuracy. P = Pronunciation. *Rounded down to .5.

After selecting the 30 video files, I then trimmed each file to a length of approximately two minutes (*M* = 2m 11s, *SD* = 14s). The rationale to reduce the length of the stimuli was primarily for practicality. The rating design of this study was fully crossed, meaning that each rater viewed and rated each speech sample. As the original videos were each up to five minutes long, watching thirty full samples would have taken the raters a minimum of 180 minutes just for the rating alone (150 minutes viewing, 30 minutes rating after the videos), plus time to enter in personal information, do the short practice session, and the final survey questions. Because this study dealt with impressions of language ability and affect without using an empirically developed rating scale, decisions could be quick and intuitive, and made on the basis of much less information than provided by a traditional rating scale. Past studies have found that impressions of affect could be made as quickly as after a 100-millisecond viewing of stimuli (Willis & Todorov, 2006), but a longer sample was necessary to form an impression of language ability. I chose two-minute segments because the test taker would have the opportunity to answer one or two test questions, thus providing my study's rater-participants a quick snapshot of the test takers' language abilities without overwhelming them with input. The two-minute mark reached approximately the mid-point of part 3 of this test section.

I trimmed each speech sample using the following procedure. Each original raw file included a set of instructions delivered by the examiner for this part of the test and an introduction to the topic, also delivered by the examiner, before each question was asked. These instructions and introductions were removed. I then trimmed endpoints for the files where there was a natural segue between test questions. That is to say, for all speech samples, the test takers had reached the end of their turn naturally before the examiner prepared the next question. I trimmed these as close to the two-minute mark as possible. In some cases, this meant trimming the end of the file before the two-minute mark (e.g., sample S04, 1m 43s) because the next question produced an extended turn that went far beyond the two-minute mark. In other cases, this meant trimming the end of the file beyond the two-minute mark (e.g., sample 24, 2m 43s), as a segue before the two-minute mark would have left an insufficient amount of speech to be rated. As noted earlier, however, most samples lasted between 1m 57s and 2m 25s.

Speech samples: Stimulated recall

Stimulated recall formed an integral part of the mixed-methods qualitative data analysis in this study. A definition, rationale, and discussion of the methods underlying stimulated recall is presented in the materials section. The stimulated recall sessions used a subset of the rating study samples as stimuli for eliciting the memories and thought processes of the raters. I used 10 of the 30 samples for the stimulated recall sessions for practical reasons. The main rationale for 10 samples was that I anticipated that the minimum number of rated samples that would provide enough data for the recall would be one third of the whole dataset, as this would allow me to choose a range of different performances, but not take too much of the stimulated recall participant's time. Any fewer than 10 would have drastically limited the range of performances available. Also, from previous experience with stimulated recall sessions based on rating two-minute samples (Burton, 2020), I anticipated that each session with 10 samples would last no more than 1.5 hours (I did not want the session to go beyond 1.5 hours), which was indeed the case. Recall sessions are generally much shorter than this amount of time, so this was already a substantial cognitive burden on the rater participants. A full recall session with all 30 files would have lasted 4.5 hours. Importantly, analyzing 10 of the 30 samples fit squarely within sequential explanatory mixed methods design principles (Creswell & Plano Clark, 2017), as I would be able to choose samples that appeared to exhibit the greatest impact of nonverbal behavior on scores in order to better target these samples for elicited responses.

Sample selection for stimulated recall. The choice of samples used in the stimulated recall was an important methodological decision for this study. It needed to be informed by the quantitative data, yet the qualitative recall sessions were inherently a subpart of the quantitative data collection itself because the stimulated recall participants completed the rating study prior to the recall sessions. I thus needed to conduct the recall sessions after a notable amount of quantitative data had been collected, yet sufficiently prior to the end of the study to ensure that enough funding was available to pay participants and so that adequate sample size estimates were reached. I settled on an analysis of 60 scores from raters in the study to determine the speech samples to use. This gave me the flexibility to recruit 20 stimulated recall participants and to

continue collecting data from non-recall participants.

The IELTS scores from Table 4.3 allowed me to rank the speech samples by proficiency level. With the IELTS scores as the baseline ranking, I was able to select stimulated recall files based on the degree to which the ranking changed between the baseline and a ranking of the samples based on the scores from the undergraduates. These selected samples would then represent the greatest deviation between official IELTS proficiency scores and undergraduate-rater ratings: I hypothesized the deviation would be due at least partially to non-linguistic criteria. To rank the samples based on scores from the undergraduates, I gathered the raw rating data from the first 60 participants having finished the study. I then conducted multi-faceted Rasch measurement (MFRM) using Facets (https://www.winsteps.com/facets.htm). I ran a partial credit model with raters, samples, and criteria (fluency, vocabulary, grammar, and comprehensibility) as facets. I then extracted the Rasch ability estimates, which I used to rank the speech samples: I set these in an Excel-like file in one column alongside the same samples' IELTS scores in a second column, and the relative ranking (the difference between the two rankings) in a third. For the stimulated recall sessions, I chose the 10 speech samples that changed ranking the most: that is, the samples that contrasted the most between the IELTS speaking score ranking and the ranking of the undergraduate-rater-determined scores calculated as Rasch ability measures. For example, Sample 29, which shared the highest IELTS score of 6.5, dropped the most of all test takers to 17th place in Rasch ability ranking, a drop of 12 points. On the other hand, Sample 13, who scored 5 on the IELTS test and was ranked in the bottom half of the test takers, rose to 24th place in the rating study according to the Rasch ability estimates. The final selected files are listed in Table 4.4. It is notable that all but two samples (S25 and S29) were in the intermediate range of 5 to 5.5 on the IELTS range. Test takers in the lower range of scores had very low rank changes within ±2 points, apart from sample S08 who dropped four places. Likewise, apart from S25 and S29, test takers in the upper range had rank changes within ±4 points.

Table 4.4
Speech Samples Selected for Stimulated Recall

Sample	Gender	IELTS score	Expressiveness	Rank change
S09	F	5	Low	+10
S12	F	5	High	+6
S13	M	5	Mid	+11
S15	F	5.5	High	+6
S16	M	5.5	Mid	+7
S17	F	5.5	High	+9
S18	F	5.5	Mid	-6
S21	F	5.5	Mid	-11
S25	F	6	High	-10
S29	F	6.5	Low	-12

Rating scales

Performance test scores can be affected by features that are not present in the rating scale (Douglas, 1994; Schoonen, 2005). Raters take into account a range of other factors, such as their own perceptions of the test takers or their perceptions of the quality of the content of the test taker's response (Kuiken & Vedder, 2014). However, empirically derived rating scales and rater training typically try to reduce this type of construct-irrelevant variance in order to measure constructs and subconstructs more precisely. In this study, the opposite was true; that is to say, I wanted to measure the amount of natural variance present when individuals watch, listen to, and make judgements about L2 speaking ability in the target language use domain. By not restricting this variance through the use of detailed rating rubrics or rater training, I hypothesized that stronger insights could be gathered about how potentially construct-relevant features such as gaze, facial expressions, posture, and gesture might impact perceptions of language ability. Knowing how this impacts non-linguist listeners may provide more generalizable information than looking at trained raters, in which case much simpler, impressionistic rating scales may be more useful.

As hypothesized in this study, individuals may use both verbal and nonverbal information to make decisions that about L2 ability. They may also use both channels of information when perceiving certain affective states and traits, such as confidence and warmth (e.g., Cuddy et al., 2011). There is some indication

that these perceptions of affect may correlate with judgements of language ability (Nagle et al., 2022). In fact, it may be the case that listener-raters' judgements of L2 ability may be directly impacted by these perceptions of affect, while nonverbal behavior itself has an indirect or even minimal effect. Thus, in order to investigate how perceptions of affect might be related to raters' scores and test takers' nonverbal behavior, I also included a short set of affect scales for the raters to use after assigning language scores. The inclusion of affect scales was also intended to prevent raters from listening only and not watching the video samples, given that language may be judged by listening alone. By including categories such as warmth and attentiveness, raters were implicitly required to watch the video as well as listen because these judgements would otherwise be harder to make.

I built a measurement instrument using semantic differentials to tap into raters' impressions of the test takers. Semantic differentials (Osgood et al., 1957; Snider & Osgood, 1969) allow researchers to gain insight about evaluations of individuals in performance settings. They are generally single word adjectives or short descriptions which are paired with their antonyms (e.g., engaged/unengaged, anxious/at ease) set on an ordinal scale 5 to10 points apart. These allow observations of a broad spectrum of attitudes and impressions, as well as the intensity and directionality of each category. Semantic differentials have a number of benefits for rating, as they are simple to understand and generally do not require training. Previously built scales (e.g., Zahn & Hopper, 1985) have largely focused on opinions and evaluations of individuals (e.g., good/bad), though semantic differential scales targeting features of speech have been used in the L2 literature as well (e.g., Burton, 2020; Harding, 2011).

I constructed 14 semantic differential scales based on both features of language and affect. Each scale contained seven points, which included a midpoint. The choice of a midpoint is largely one of stance rather than psychometric properties; without a midpoint, raters must choose a direction of an effect (anxious or not anxious), and with a midpoint it is possible to see someone as relatively neutral. I chose a scale with seven points because this is the smallest scale with a midpoint with desirable psychometric properties, as scales with 5 or fewer points may have attenuated precision (Simms et al., 2019). Language features on the scale included fluency, vocabulary, grammar, and comprehensibility. Comprehensibility was chosen rather

than pronunciation because there is ongoing work on comprehensibility in this area (e.g., Nagle et al., 2022; Tsunemoto et al., 2022). I also believed the novice raters would understand comprehensibility better than pronunciation, as constituent parts of skillful pronunciation (e.g., phonemic control, appropriate stress) may be unfamiliar to linguistic laypeople, and they may have resorted to judgements of accent. I chose affect measures based on the L2 literature, selecting those that have a confirmed relationship with proficiency, including engagement (Ducasse & Brown, 2009; Jenkins & Parra, 2003; May, 2009; Nakatsuhara et al., 2021a; Sato & McNamara, 2019), anxiety (Sato & McNamara, 2019; Thompson, 2016), confidence (Ducasse & Brown, 2009; May, 2009, 2011; Thompson, 2016), and expressiveness/happiness (Jenkins & Parra, 2003; Thompson, 2016). I also chose two measures related to engagement that reflect the socio-affective orientations of the test taker: attentiveness (Ducasse & Brown, 2009; May, 2011) and interactiveness (related to interactional competence; Galaczi & Taylor, 2018; Plough et al., 2018), as these stances may also factor into raters' judgements. I included warmth, attitude, and competence (Cuddy et al., 2011) as these traits were found to relate to positive or negative outcomes in organizational psychology and could possibly relate to outcomes in this study as well. The scale was piloted with 25 participants (students in the target demographic, enrolled in a teacher education course) and analyzed using Facets. It was found to function as intended, with scale units ordered as intended, meaningful separability amongst the seven scale units, and no misfitting scales overall. Pilot participants indicated that the scales were simple and intuitive to use. The full scale is provided in Figure 4.4.

The scales were presented in the online survey system in the following way. First, the polarity of the adjectives alternated in the operational scales so that for some scales the positive adjective was on the left, while for other scales the negative adjective was placed on the left. This was to prevent *survey acquiescence bias* (Iwaniec, 2019), such as marking all positive answers (7s, in this study) for all categories, as raters would have to carefully read each scale to know whether a 1 or a 7 was positive or negative. Second, the language categories were always presented first in the same order as Figure 4.4, followed by a randomized order in the affect categories for each speech sample. Randomizing the order was meant to reduce order/primacy/recency effects of the various affect scales. Primacy was desirable for language

features, as I wanted raters to primarily focus on paying attention to the features of L2 ability.

Figure 4.4
Language Features and Affect Rating Scales

Rate the **speaker's language** on the following elements:		
Fluent	__:__:__:__:__:__:__	*Disfluent*
Strong vocabulary	__:__:__:__:__:__:__	*Weak vocabulary*
Strong grammar	__:__:__:__:__:__:__	*Weak grammar*
Comprehensible	__:__:__:__:__:__:__	*Incomprehensible*
Rate the **speaker** on the following elements:		
Engaged	__:__:__:__:__:__:__	*Disengaged*
At Ease	__:__:__:__:__:__:__	*Anxious*
Confident	__:__:__:__:__:__:__	*Not confident*
Warm	__:__:__:__:__:__:__	*Cold*
Attentive	__:__:__:__:__:__:__	*Inattentive*
Expressive	__:__:__:__:__:__:__	*Inexpressive*
Happy	__:__:__:__:__:__:__	*Unhappy*
Competent	__:__:__:__:__:__:__	*Incompetent*
Interactive	__:__:__:__:__:__:__	*Non-interactive*
Positive attitude	__:__:__:__:__:__:__	*Negative attitude*

Sign-up survey

Prior to being invited to take part in the study, individuals interested in participating completed a questionnaire which included a range of demographic variables used to determine their eligibility. The questionnaire had fields for the participants' names (used to address participants in automated e-mails and piped text in the survey), e-mail address, year of birth, gender, nationality, L2 status, L1, year of study at Michigan State University, and major. There were additionally questions asking whether participants had access to a quiet, distraction-free space to complete the survey and whether participants would be willing to take part in the stimulated recall. The survey text is presented in Appendix B.

Online rating platform

The online rating platform was constructed and hosted in Qualtrics. I created two-formats for the survey: a two-day format used with the majority of the raters, and a one-day format used with the stimulated

recall participants. The rationale for the two-day design, which spread rating out in equal halves on two days spaced out by 24 hours, was to reduce rater fatigue. The novice raters were unaccustomed to rating speech samples, and rater fatigue can reduce the quality of ratings. The one-day rating design was used with the stimulated recall participants to ensure that the rating of all samples was fresh and recent in the raters' minds when they conducted the stimulated recall instead of the longer two-day design. Both designs contained the same components: an introduction, instructions and practice, the speech samples and rating scales, and the follow-up questionnaire. The follow-up questionnaire was only included in the second day of the two-day format. Images of the instructions and practice for the study are presented in Appendix B. These images mirror how the study looked overall, with the exception that feedback was not presented in live rating. I am not able to share direct links to the survey as it includes proprietary information from IELTS protected by a non-disclosure agreement.

Introduction. The introduction to the online rating platform included a description of the study, instructions on how to set up for the study, data verification, and consent/non-disclosure agreements (Appendix A). The instructions asked participants to secure time and a quiet place to conduct the study in one sitting, and to use headphones if available. Information was given to participants on follow-up activity after the study (e.g., stimulated recall, compensation), as well as contact information in case problems arose. This section also verified participants' names and e-mail addresses using piped data from the original survey sign-up, in Appendix B. Finally, participants were asked to agree to both the consent form and the non-disclosure agreement separately. Any response of "no" to either the consent or non-disclosure agreement closed the survey.

Introductions and practice. The introduction section introduced the task for participants and the terms used in the rating scales. Terms were not explicitly defined because it was desirable for participants to bring their own internal definitions of the terms to the rating scenario. While this can have a negative impact on the reliability of scores, this was advantageous because in the target language use domain, individuals make judgements about others using their own internalized understanding of the world around them. Providing extended definitions may have introduced unnecessary confusion or difficulty with the

task if raters were unfamiliar with certain ideas. For example, defining interactiveness as showing evidence of interactional competence may have confused raters unnecessarily, as they would not be accustomed to thinking about interactional skills as defined in applied linguistics.

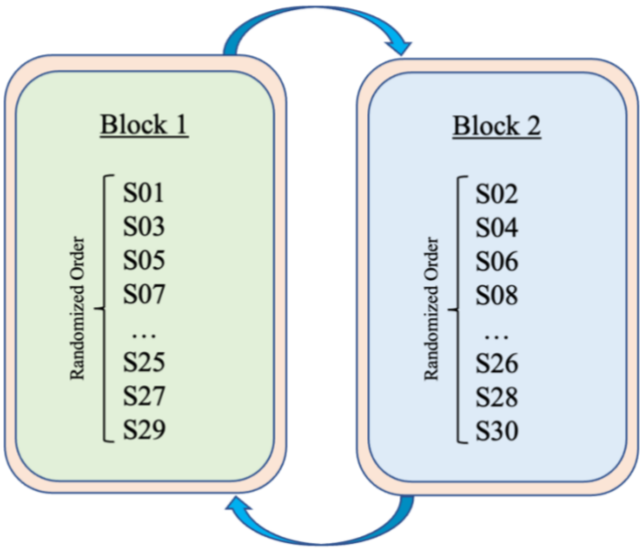
The practice section immediately followed the introduction. Each participant viewed the same two cases, one of a higher ability speaker and one of a lower ability speaker. The practice videos were drawn from samples I had recorded in previous studies with consent from speakers. Participants were encouraged to consider the performance and rate the sample on the 14 scales. After rating, which was not scored, raters received minimal feedback about the performance (but not their ratings). For example, the feedback after the first performance was: "Although the speaker struggles to understand the question at first, overall her language is fairly strong once she begins speaking. She manages to communicate fairly effectively." Explicit benchmarked scores were not provided to the raters, as it was hoped that raters would develop their own internal definitions of the scale categories. Both the introduction and practice materials are available in Appendix B. The introduction and practice sections were available to participants on both days of the study for the two-day format, but participants were able to skip the practice on the second day if they desired.

Speech samples and rating scales. After the practice session concluded, participants were directed to the rating portion of the study. The speech samples were presented in a random (but even) order for all participants, meaning that all participants rated all samples one time, but each participant was presented samples in a different order. The speech sample stimuli were presented on individual pages without the presence of rating scales. Rating scales were presented only after the videos had finished and were on a separate page. The samples and scales were divided on separate pages to reduce distractions and to encourage participants to watch the video the entire time it was playing, as otherwise the raters may have looked through the rating scales while listening. I embedded Java code so that the samples could be played only one time with no pausing, no other video controls, and also no ability to download the files. I also entered Java code so that the videos would be presented in the maximum size possible (rather than the smaller default versions) within Qualtrics in their internet browser. I encoded a large video size so that participants would have a much larger visual area to pay attention to. The videos would have taken up the

majority of the participants' browsers. Unfortunately, it is impossible to estimate the exact dimensions of these for each person because screen sizes were ultimately different, and each person could control their browser window size. As mentioned earlier, the rating scales following the samples were presented with language features first in the same order (fluency, vocabulary, grammar, and comprehensibility), while affect scales were presented in a random order for each participant.

Participants completing the two-day study and the one-day study had slightly different rating designs. For the two-day study, I divided the 30 samples into blocks of 15 by odd even numbers. Dividing by odds or evens ensured the same distribution of proficiency levels for each day. The two blocks were counterbalanced by randomly assigning participants one of the two blocks to begin with on the first day of the study. Twenty-four hours after completing the first day of the study, participants received an automated e-mail from Qualtrics giving access to the remaining block of questions. This design is shown in Figure 4.5. Participants completing the one-day study were provided all 30 speech samples presented randomly. After finishing 15 samples, however, a break screen was presented to encourage the raters to rest for a moment in order to reduce rater fatigue.

Figure 4.5
Counterbalanced Blocks for Day 1 and Day 2 of the Two-Day Study



Follow-up survey. After completing the online rating study, participants were directed to an

optional brief questionnaire about their experience while participating in the study. The follow-up questions were used to monitor ratings as they were submitted during the study to ensure that participants were not experiencing technical issues. The questions also served as a way to verify how raters felt about the scales themselves and allowed them to comment on any doubts or difficulties while rating. A final section in the questionnaire elicited information regarding how much the raters felt that various facets in the testing situation impacted their ratings of language and affect separately. The follow-up survey items are presented in Appendix C but were not analyzed in this dissertation.

Stimulated verbal recall

Second language research has traditionally relied on *product data*, such as the rating scores in this study, to explain phenomena relating to language development. Product data, however, give limited insight into the nature of *how* and *why* individuals perform in particular ways, and thus *process data* may support investigations of human behavior through the triangulation of the two data sources. While retrospective interviews, questionnaires, and other forms of qualitative data collection may be used to investigate the process data of the internal working of cognition while performing tasks, one of the more popular research methods over the past 40 years has been the use of verbal protocols. Verbal protocols, arising from research in cognitive psychology, seek "to understand in detail the mechanisms and internal structure of cognitive processes" (Ericsson & Simon, 1993, p.1). They may occur as concurrent think-aloud protocols, which take place while a task is performed, or they may take place after performing a task as retrospective or stimulated verbal protocols. Because the focus in this study is on the rating of speech, concurrent protocols are unmanageable because participants' verbalizations of their thought processes would overlap with the speech being heard. In addition, retrospective recalls, which take place after an entire task is performed, are not ideal because the task in this study was to listen to a series of speech samples, and human memory is limited in what it can store and reproduce. For this reason I chose to use stimulated verbal recalls for use in this study.

Stimulated verbal recall (or stimulated recall, for short), aim to enable participants "to relive an original situation with vividness and accuracy if he is presented with a large number of the cues or stimuli

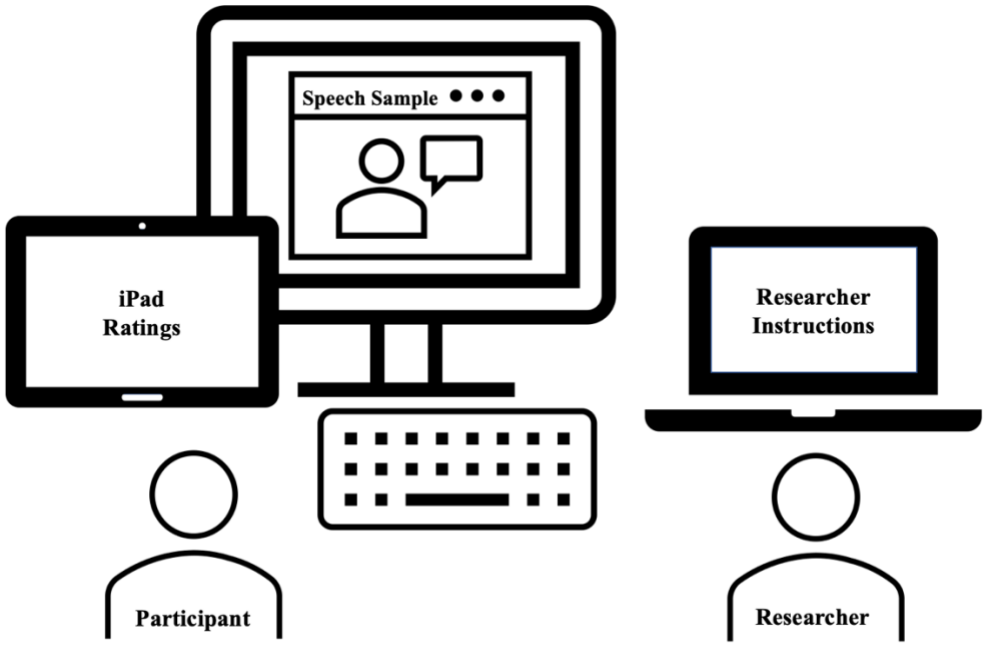
which occurred during the original situation" (Bloom, 1953, p. 161). This type of method generally uses recordings of a task in progress to stimulate the memory of participants, which they are then encouraged to verbalize by pausing the task recordings when memories occur. Stimulated recall designs are useful for the analysis of rating process data because they can be relatively non-intrusive and intuitive to perform, do not have problems of reactivity (negatively impacting scoring data, which has already been collected), and are generally considered to be veridical (revealing the true nature of cognitive process while rating rather than other aspects of cognition) as long as they follow strict guidelines (Bowles, 2018; Egi, 2008). There is a well-established tradition of the use of stimulated recall in L2 research in both SLA and language testing with guidelines for best practices (Gass & Mackey, 2016), which I used in the design of this study.

The stimulated recall sessions in this study formed an integral part of the sequential explanatory design I adopted, as the speech samples in the recall had been identified as exhibiting scoring patterns that merited further study. The method that I previously described involved identifying samples which were scored much higher or lower than the original ranking of the IELTS scores, indicating a potential effect of behavior on the resulting scores. Ten of the 30 videos showing the largest differences in the ranking were chosen for this study.

Each session included carefully drafted instructions, which were piloted operationally with one participant. I drafted the instructions using examples from Gass and Mackey (2016), making sure to emphasize that participants were to recall their memories from the rating session and not their thoughts at the time of rating. Participants were not asked to speak about any particular rating categories, nor were they limited to discussing only verbal aspects of speech in the samples. Participants were shown an example of how to operate the recall session using the space bar of an iMac computer to pause and start the videos. There was no practice session as the recall sessions are generally intuitive, and I wanted to limit any unnecessary time spent due to the number of videos to be watched. Participants were provided with an iPad next to the iMac computer that displayed their score reports of the 10 samples obtained from Qualtrics. That is to say, the pdf printout from Qualtrics showed the same format of the rating scales, which I hoped would help stimulate the raters' memories further. I sat to the right of each participant with instructions and score

results available on my laptop computer. The setup is displayed in Figure 4.6. I drafted a set of probe questions to be used when pausing the sample videos. The session finalized with a semi-structured interview to target specific aspects of my research questions, as well as to follow up with particular comments made by the participants during the session. The drafted instructions, probe questions, and interview questions are available in Appendix D.

Figure 4.6
Stimulated Recall Setup



As explained in the above sections, I used ongoing rating data to identify a subset of 10 samples to be analyzed using stimulated recall. The videos were presented in a random order for each participant. I invited a subset of 20 raters to take part in these sessions. Each recall session lasted an average of one hour and 11 minutes (*SD* = 18m 58s). The resulting dataset included 200 unique recalls, 20 follow-up interviews, and 23 hours and 53 minutes of data. The transcribed dataset, including both stimulated recall and interview content, contained 157,894 words.

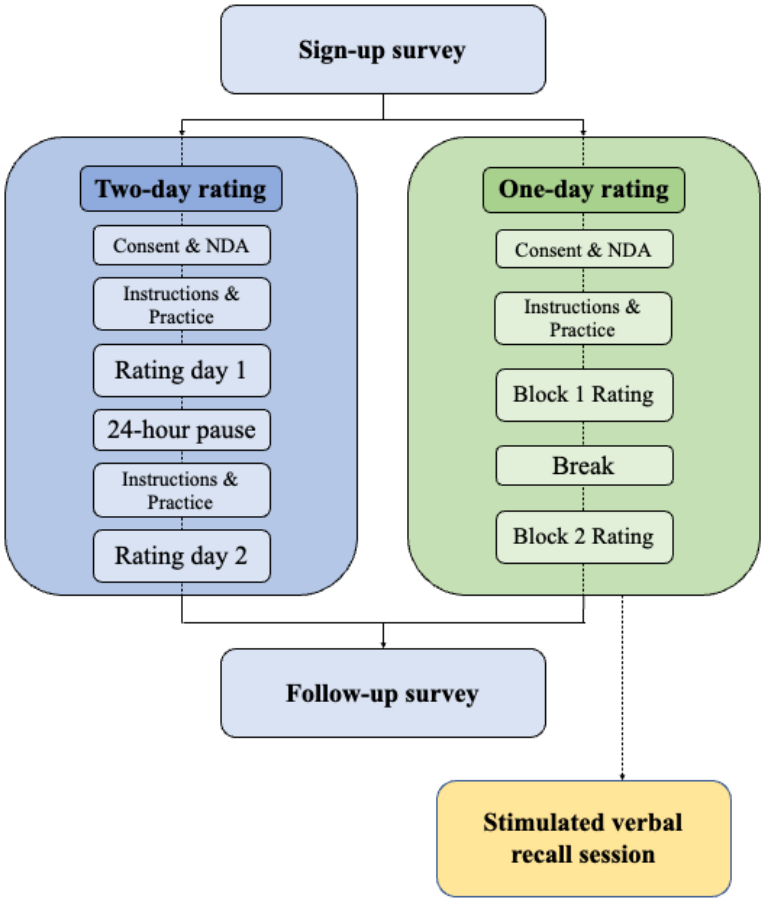
Procedures

Rating study procedures

The rating study procedures are outlined in Figure 4.7. Data collection took place between

December 2021 and April 2022. Participants first indicated their interest in the study by completing the sign-up survey in Appendix B, which collected demographic information used to identify eligible participants. Following this, I invited participants in batches of 10 to 50 individuals to participate in the two-day rating design. Participants were sent e-mail invitations through Qualtrics that piped their unique participant IDs, names, and e-mails into the rating flow. Twenty-four hours after completing day 1 of the study, participants were automatically notified that day 2 of the study was available. Participants were not obliged to complete day 2 of the study immediately but could choose a day and time of their preference. Attrition (participants that did not complete day 2 of the study) was low, at 8%. The follow-up survey was included at the end of the day 2 survey. Upon completion of this study, participants were provided a $30 Amazon e-gift card.

Figure 4.7
Data Collection Procedures



After identifying samples to be targeted for the stimulated recall session, I invited participants in batches of 5 to10 individuals to participate in the one-day rating plus stimulated recall sessions. Those that indicated interest in completing both the rating and stimulated recall sessions were sent links to choose a date and time for their recall sessions, after which they received instructions on how to proceed. Three days before the stimulated recall, I sent each participant detailed instructions outlining each step of the process to ensure that the rating was completed less than 24 hours before the stimulated recall session. One day before the recall session participants received a link to the rating study. When the study was complete, I then sent each participant directions to the laboratory used for this session. All e-mail communications to participants for both the two-day rating and one-day rating designs are included in Appendix E.

Stimulated verbal recall session procedures

The stimulated recall sessions took place in March and April, 2022. The sessions took place in person at the SLA Knowledge and Production Lab in Wells Hall. All participants had completed the rating session no more than 24 hours before the recall session began. Participants were provided with a bottle of water and invited to sit at a Mac terminal. I sat to the right of the participants and used my laptop to conduct the session. I welcomed each participant with small talk to put them at ease, and then explained the content of the session. Participants then signed a second consent form agreeing to the audio recording of the session, indicated at the end of Appendix A. I then gave the participants instructions on the stimulated recall session and demonstrated how to start and stop each video. I provided participants with an iPad that showed the scores they awarded for each speech sample. I then began the audio recording by both recording a screen share on the Mac computer (in order to capture time stamps of the video when the participant paused) and using an external digital recording device. I then began the session. Participants watched all 10 speech samples and recalled their thought processes, which lasted an average of one hour and 11 minutes (*SD* = 18m 58s). I then debriefed each participant in a semi-structured interview. Each stimulated recall participant was compensated with a $50 Amazon e-gift card. Stimulated recall instructions, probe questions, and interview questions are available in Appendix D.

Analysis

Software

iMotions. I analyzed the video speech samples using iMotions software (Version 9.0; iMotions, 2017). I extracted three indices for analysis in this study: engagement, valence, and attention. I chose these three measures because each captured complex combinations of facial movements that related to features raters identified in the literature (e.g., expressiveness, positivity, and gaze direction). Affectiva (n.d.) (the parent company that produces the facial algorithm Affdex for the software iMotions) provided descriptions of how each measure is compiled. Engagement is defined by iMotions' facial recognition algorithm Affdex as a measure of overall expressiveness derived from the participant's facial muscle activation. Facial muscles contributing to this measure include eyebrow raising and furrowing, cheek raising, nose wrinkling, mouth movements, and chin raising. Engagement is thus not an indication of positive or negative emotion but rather an indication of how non-neutral a participant may appear. Valence, on the other hand, is defined by Affdex as a measure of the positive or negative emotions exhibited by the participant. Valence is calculated by smiling and cheek raising for the positive end of the scale and brow raising/furrowing, nose wrinkling, mouth frowning, lip pressing, and chin raising for the negative end. Attention is a measure of gaze and head turns directed towards the stimulus source (in this case, an examiner visible on a laptop computer with a camera embedded in the upper bezel). Data output for these action units/expression metrics, emotions and states is in the form of a probability-based confidence score of 0 to 100 for each video frame. For example, if a frame receives a value of 87 on engagement, the algorithm has classified this instance as highly likely that the individual is classified as engaged. Valence, or the strength of positive or negative emotions, is the only measure that uses a scale from -100 to 100, where -100 is a highly probable negative overall response and 100 is a highly probably positive response. The final output of each response is a table of probability measures for each frame of video analyzed.

NVivo. I used the qualitative data analysis software NVivo for Mac (Version 12; qsrinternational.com) to analyze the stimulated recall data. NVivo is a popular tool used to organize qualitative data by using cases, which in this dataset were two kinds: the stimulated recall (a) raters and (b)

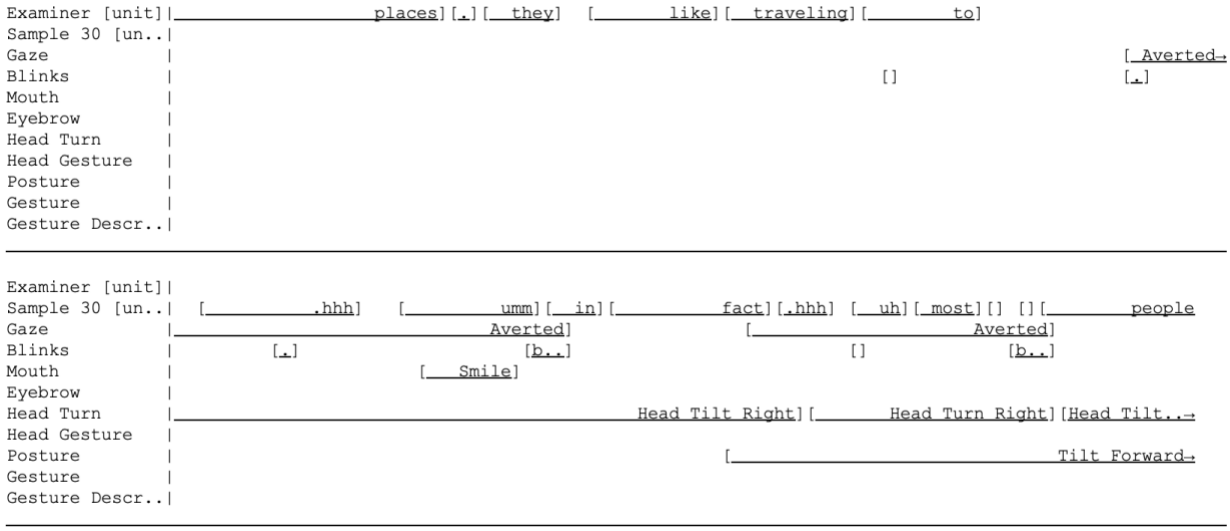
speech samples; and nodes, which are annotations of themes that arose in the dataset. The power of NVivo is that once the dataset is annotated, it allows researchers to analyze frequencies and cross instances of nodes, such as fluency and eye gaze, to find deeper patterns in the dataset.

ELAN. I used ELAN (Max Planck Institute, 2020) to annotate both speech and nonverbal behavior in the 30 speaking test files. For this study, I drew on an annotation system that I developed for the study of nonverbal behavior used in repair sequences (Burton, 2021a). This system included four tiers for verbal information (simplified conversation analysis and individual words) plus seven tiers of nonverbal behavior (gaze, blinks, mouth movements, eyebrow movements, head position, posture, and gesture). For this dissertation study I refined and added to this annotation scheme as behaviors became salient in the example performances. I added a tier for head gestures (nods and shakes), and I added in additional behaviors to the system such as rocking back and forth, shifting posture, and shifting eye gaze. I also added an additional category for occasional behaviors that were otherwise uncategorized, such as shoulder shrugs and swallowing. The full annotation scheme is provided in Appendix F.

ELAN produced two types of data useful for the analysis of stimulated recall data. These were used to compare the spoken transcripts from the raters with evidence of what test takers did during the samples in reality. One data type was the multimodal transcripts. These were fine-grained, tier-by-tier descriptions of the unfolding interaction between the examiner and test taker. One example, taken from sample 30, is shown in Figure 4.8. The top line in this sample is the annotation word-by-word for the examiner, who is finishing his question. The next line is the test taker in sample 30, likewise annotated word-by-word, including breath marks *.hhh* and filled pauses *umm*. Next, there are nine tiers of nonverbal behavior, not all of which are annotated because not all behaviors appeared in this excerpt. Gaze was only annotated when averted or shifting, and mutual gaze was left uncoded. In this sample, the test taker averted her gaze when she began her turn, reestablishing mutual gaze with the examiner at the word *fact*. She smiled during her filled pause and blinked six times. She began tilting her head right as she began her turn, first to the right, followed by a full turn right. Her posture began leaning forward at the word *fact*. There were no examples of eyebrow movement, head gestures (e.g., nods), or gestures in this excerpt. In the analysis, I removed

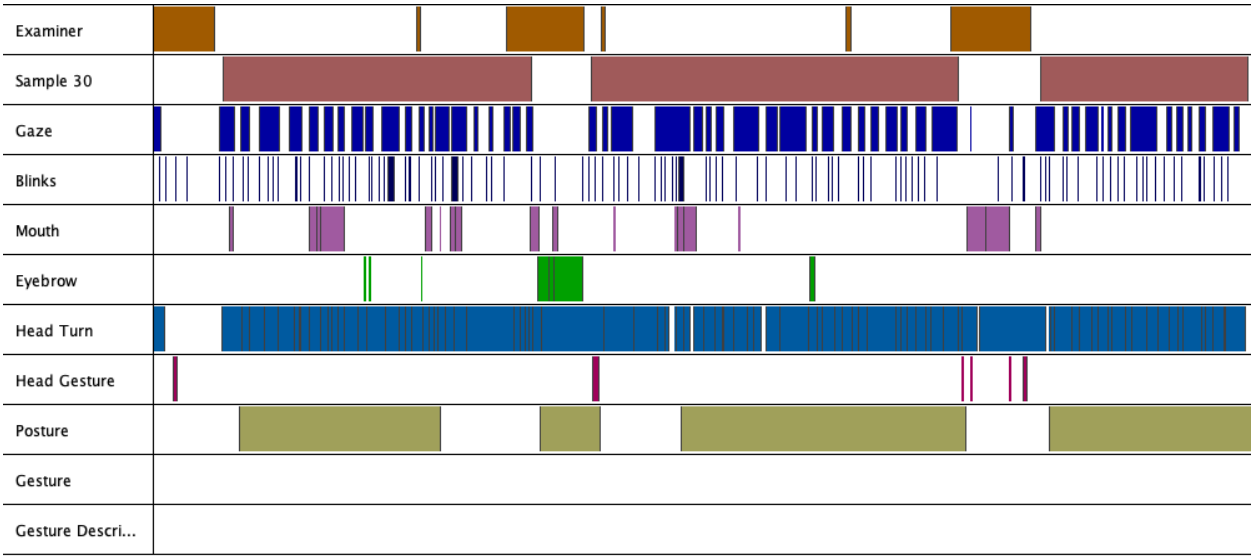
empty tiers to shorten the transcripts and make them more readable, as appropriate.

Figure 4.8
Multimodal Transcript from ELAN (Sample 30)



The second data type that ELAN produces is the annotation density graph. These graphs show the appearance of behaviors throughout the entire two-minute sample and are useful when considering the sample behavior as a whole. The annotation density plot for sample 30 is shown in Figure 4.9. This plot is arranged similarly to the multimodal transcript, but solid bars represent entire moments when behaviors were coded within that tier. In this example, the test taker spends most of the time talking, with the examiner asking questions or backchanneling six times. The test taker used variable gaze patterns, withdrawing and reestablishing gaze frequently. She did not sustain smiles for long periods of time, but she did smile frequently. She used few eyebrow movements, though notably moving them in a sustained manner during the examiner's second question. She turned her head frequently during this sample, rarely keeping a neutral position. She nodded or shook her head (head gestures) six or seven times, particularly during the examiner's questions. She changed her posture frequently as well and produced no visible gestures.

Figure 4.9
Annotation Density Plot from ELAN (Sample 30)



Data Preparation

Rating data. The rating data for the 100 participants were extracted from Qualtrics. I prepared the dataset for analysis using R (R Core Team, 2022). Because the scales were presented to participants with alternating scale polarity (1 and 7 represented both positive and negative trait judgements depending on the scale), I adjusted the polarity of all scales such that negative judgements (e.g., weak vocabulary, anxious, cold) were reordered with 1 being the lowest endpoint, and positive judgements (e.g., strong vocabulary, at ease, warm) were reordered such that 7 was the highest endpoint. The final dataset included the variables listed in Appendix G.

iMotions. Raw iMotions data were prepared prior to use in the study. Raw scores "represent the classification results of the facial expression engine for a certain respondent compared to the facial expressions stored in the global database" (iMotions, 2017, p. 31). Baseline corrections were not applied to these samples, as these would center all participants identically, thus blurring the meaning of neutral states (one participant's neutral state may be highly aroused while another may appear more neutral). This correction is recommended by iMotions (2017) for the investigation of individuals' *relative* changes in their own behavior, but in this study the focus is to contrast differences between individuals, not within. I also

avoided thresholding output indices on either time or amplitude. Thresholding is important when determining the likelihood of the appearance of an individual feature such as a smile or brow furrow (such as the boundaries of fixations in eye-tracking). These boundaries may be determined by the strength of the facial muscles when producing the behavior (amplitude) or the amount of time the behavior is held (time). Thresholding allows researchers to investigate longer-lasting emotions or emotions that exceed a certain strength (iMotions, 2017). However, in the context of this study, it is unknown whether small variations in change may affect the overall perception of test takers, or whether rapid bursts are noticeable, both of which could be eliminated if time or amplitude thresholding were implemented. For this reason, I used the raw probabilistic data in this study. Raw scores are most appropriate when comparing different individuals or groups of respondents (iMotions, 2017).

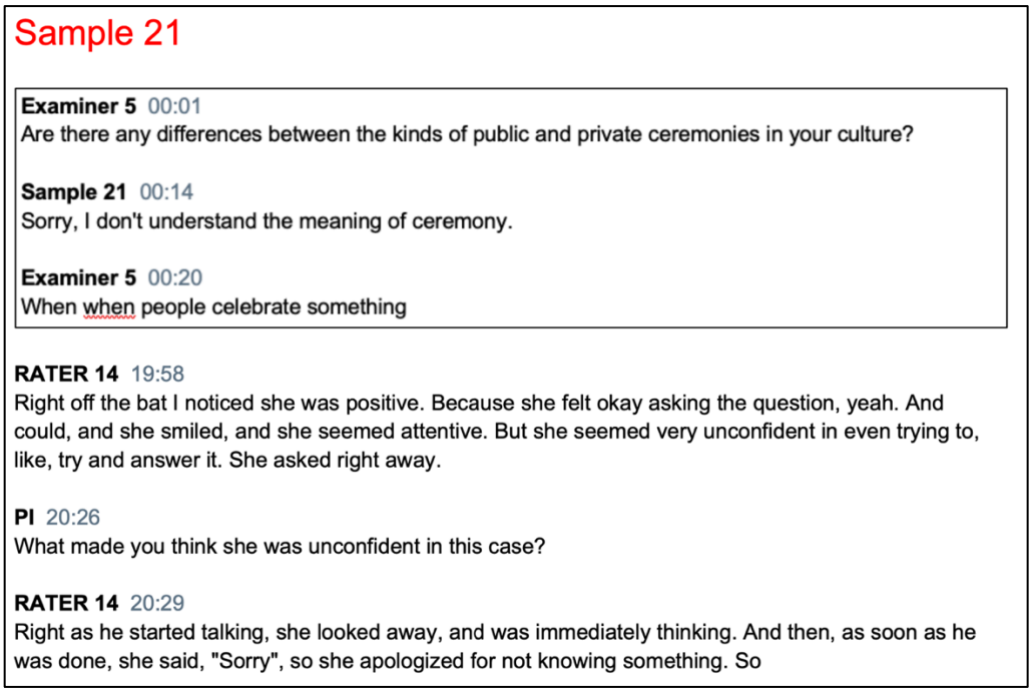
iMotions extracted 117,221 frames of data from the 30 speech samples at 30 frames per second, totaling 66m of video. I averaged engagement, valence, and attention values for each speech sample for analysis. The averages represented the average of the confidence measures for each value, in other words, the overall strength of engagement, valence, or attention for each speech sample across the entire video. However, the values across the videos were constantly in flux as the test takers' behavior shifted. For this reason, I also computed standard deviations of each value to represent the amount of standardized variance for each test taker, as this would capture the number of changes in behavior, which may also be meaningful in analyses.

Stimulated recall transcripts. I prepared the stimulated recall audio recordings for analysis by first conducting an automated transcription of the individual speech samples in Otter.ai (https://otter.ai), which I then listened to individually and corrected. Otter.ai automatically transcribes audio files and automatically recognizes speakers throughout the transcripts. It includes speaker names, timestamps, and the text produced. These are transcribed orthographically without the inclusion of paralinguistic features or filled pauses, except where these were extended. Where words were undetectable or difficult to understand, I indicated these with brackets [Unknown] or [???]. Otter.ai provided commas where natural pauses were

in the recordings.

After the speech samples were fully transcribed, I uploaded the stimulated recall audio recordings to Otter.ai and repeated the same process. In the final transcript, I included the text from the original speech samples up to the point the recordings were paused in the stimulated recall rather than the entire block of text. A sample of the transcript is provided in Figure 4.10. The speech sample is labeled at the top, in this case Sample 21. The transcribed sample text is shown in the text box, where Examiner 5 and the test taker in Sample 21 are talking. The timestamps in the textbox refer to the times the text in the speech sample were produced. Following this, Rater 14, the 14th stimulated recall participant, provided their recall, with a separate timestamp that refers to the stimulated recall session. Following their comment, I extended the recall as PI (principal investigator) with a standardized question. Rater 14 then answers the question, provided an extended recall, and then continued the audio. Note that when Rater 14 referred to an explicit quotation of text from the test taker, I used quotation marks in the transcript to identify this word. Once these transcripts were complete, I uploaded them to NVivo.

Figure 4.10
Stimulated Recall Transcript



NVivo. Data preparation in NVivo involved uploading the transcripts for the stimulated recall

participants. Each participant's file received a case coding of the participant number, and within these files each discussion of a speech sample received a case coding of the sample number. This resulted in 20 cases for rater participants (1–20) and 10 cases for speech samples. Final debriefing interviews were coded as a third case, but these data are substantially different from stimulated recall data and beyond the scope of analysis for this dissertation.

ELAN. Multimodal speech and behavior transcriptions for the 30 speech samples were conducted in ELAN. Using the Otter.ai transcriptions, I added in speech tiers to the dataset. The only difference in transcription style was that I transcribed filled pauses, audible inhale/exhale, and laughing for each sample. I also annotated a tier of the production of each individual word. Because the behavioral data were to serve for analysis with the stimulated recall data (and possible subsequent analyses post-dissertation), it was paramount to ensure that the transcripts with behavioral data were as accurate and precise as possible. At the same time, dense multimodal transcription is extremely time consuming. I hired two research assistants (Master of Arts graduate students in Teaching English to Speakers of Other Languages (TESOL)) to assist with the behavioral annotations. Instead of a standard 20% reliability check, I designed an iterative training course covering six samples (20% of the dataset) to ensure that the research assistants and I annotated the samples as similarly as possible prior to annotating the entire dataset. During these training sessions, additional behaviors became salient (e.g., throat movements during gulping, shoulder shrugging, tongue sticking out), and I updated the annotation scheme to include these behaviors.

The first stage of training included practice, feedback, and consensus. I trained the research assistants how to use ELAN for multimodal transcription and annotation. I showed them how to annotate each line of behavior, providing information on how to define boundaries for each. For this stage, we worked with two samples, S01 and S16. I annotated both samples fully, and on separate days the research assistants each submitted their annotations for individual tiers of behavior. In other words, these feedback sessions focused individually on gaze, blinking, eyebrow movements, etc. In this round of feedback, I compared my annotations with the assistants' two annotations, and we discussed agreements and disagreements, coming to a consensus on each tier of these files.

In the second stage of training, each research assistant annotated a different set of two files (06 and 21; 11 and 26), and I annotated all four. The assistants completed all tiers of behavioral annotation, and then I gave feedback to each assistant after comparing my annotations with theirs. I calculated reliability separately for number of annotations per tier (e.g., gaze, blinks, posture) and duration of annotations per tier. This was because the number of discrete annotations could be misleading, as each annotation could last 1 millisecond up to several seconds. The average Pearson correlations between each research assistant and me on the total number of annotations was .95 and the duration of annotations was .91, which I considered to be a strong measure of agreement. For each file, I offered feedback on both our agreements and disagreements. For each annotation in which we did not align, we reached consensus following feedback and analysis.

Following training, the research assistants independently annotated the remaining 24 speech samples. One research assistant annotated 8 files, and the second annotated 16. The numbers differed due to the assistants' differing work-availability times. After annotation, the research assistant who annotated 16 of the files and I reviewed all files. Multimodal transcripts were then compiled in ELAN for use in the stimulated recall analysis.

Data Analysis

Rating data cleaning. All variables used in this study are listed in Appendix G. The first part of the statistical analysis involved ensuring the dataset was sufficiently reliable for analysis. Because the study involved the use of novice raters that completed the study completely online for a gift certificate, it was anticipated that some individuals would be careless in their ratings or directly negligent. It was thus necessary to inspect the dataset for undesirable responses. I cleaned the dataset using a combination of methods. First, I calculated Spearman correlations for each of the 99 raters to determine whether they agreed with the rest of the group on stronger and weaker performances overall. For each rater, I calculated the average of each raters' four language outcome correlations with all other 98 raters. I then calculated a grand mean of all raters' average correlations to produce a rough estimate of their overall agreement with the

group. I flagged raters with correlations lower than .4.

Second, I ran MFRM analysis—with 99 raters, 30 speech samples, and four language criterion measures as facets in a partial credit model—to determine misfitting raters using fit statistics. I only used language criterion measures of fluency, vocabulary, grammar, and comprehensibility in this model, as there is an expectation of unidimensionality in Rasch models (Aryadoust et al., 2021), and I posited that including affect scales would result in multidimensionality. I used Bond et al.'s (2020) and Linacre's (2002) criterion of fitness between .5 and 1.5 of infit mean square values, and Wright and Linacre's (1994) criterion of a maximum of 2.0 for outfit means square values. Infit statistics report the mean squared residuals of raters on inlying performances, while outfit statistics identify outlying performances (Linacre, 2002).

Finally, as outliers can have a negative impact on ordinal regression (Tabachnik & Fidell, 2013), I checked the dataset for patterns in multivariate outliers amongst participants. I used the four language ability scales to check for outliers in the four-dimensional data. I used Mahalanobis distance to detect values that veered from a normal distribution in a Q-Q plot. These values were individual scores on language criterion measures. I then produced a frequency table of raters that produced the highest number of multivariate outliers to be flagged for possible erratic behavior.

With these three methods (low rank correlations with the overall group, Rasch misfit, and multivariate outliers), I removed 16 raters whose scores did not align with the majority of the group.

Integrity of rating data. The second part of the analysis involved providing evidence that the rating data could be used to produce meaningful inferences. I calculated descriptive statistics to show overall trends in the dataset. I report Cronbach's alpha for each scale and the distributions of rating scale scores. I also report scale, sample, and rater data derived from the partial credit Rasch analysis (only using language criterion scales due to dimensionality concerns) to provide evidence that the scales, raters, and samples functioned within acceptable parameters. In addition, I report intraclass correlation coefficients (ICCs) as a measure of interrater reliability. ICCs provide a measure of both correlation and degree of agreement, taking into account the mixed effects nature of the dataset (raters x samples). ICCs below .5 are

low, between .5 and .75 are moderate, .75 and .9 are good, and above .9 are high (Koo & Li, 2016).

Affect analysis. To answer the first research question, I began by calculating inter-scale correlations to determine the associations amongst all scale categories. I used polychoric correlations because Pearson correlations are likely to attenuate relationships amongst ordinal or ordered categorical variables (Winke et al., 2022). I used the *polychoric* function in the *psych* package (version 2.0.8) in R. In order to produce inferential data about the relationship between affect and language proficiency, it was desirable to reduce the dataset to use a smaller number of components for analysis. Although principal components analysis is mathematically more appropriate for data reduction and summarization when there are no theoretical relationships underlying variables (Tabachnik & Fidell, 2013), the variables in this study were semantically related, and I hypothesized that an underlying factor structure existed that would enhance interpretability after data reduction (rather than a component structure that cannot be interpreted). Given that interpretability was a key desire in this study, I checked the factor structure underlying the 10 affect scales. Because violating assumptions such as multicollinearity can result in unstable factor scores, I checked assumptions for factorability to make sure the factor solution would be robust. Afterwards, I reduced the dataset using exploratory factor analysis with the polychoric correlation matrix using maximum likelihood estimation and a promax rotation. I then used the factor scores from factor analysis to conduct mixed-effects ordinal regression, as described in detail below, to observe relationships between the factors and language proficiency outcomes.

iMotions behavioral data. Descriptive information on means and standard deviations is presented for each of the iMotions variables: valence, engagement, and attention. In addition, graphical data exploring the distribution of these variables is presented for descriptive analysis. I then built models using the means of the iMotions variables as predictors using mixed-effects ordinal regression, as described below, to answer research question 2.1. To answer research question 2.2, I used the scaled scores from IELTS as interaction terms in the same models. Throughout the dissertation, these scores will be referred to as *base proficiency scores*, as they were external and measured prior to the research taking place. The analyses described above were confirmatory and preregistered in the Open Science Framework (https://osf.io/u6243).

An exploratory analysis, not hypothesized a priori, involved using a separate set of predictors using the standard deviations of the variables rather than the means. Using standard deviations allowed testing whether variance in behavior (instead of the average level) would predict proficiency measures.

Ordinal regression. I used cumulative logit mixed effects models—a proportional odds model with ordinal outcomes—to test for relationships amongst the extracted factor scores and the language variables, as well as the relationships amongst the iMotions behavioral indices and the language variables. I used the *clmm* function in the *ordinal* package (v. 2019.12–10) in R to account for mixed effects of raters and samples.

Most assumptions for the four models were met: the dependent variable was ordinal, the independent variables were continuous (factor scores and iMotions scores), and there was no evidence of multicollinearity. The fourth assumption, that of proportional odds, or parallel regression slopes, was checked using Brant's tests (Brant, 1990) using the *brant* package (v. 0.3–0) in R on models with random effects removed (*polr* models). This is due to the fact that there are currently no packages that are able to check for parallel regression slopes using *clmm*, and current recommendations suggest using base models without random effects as sufficient evidence of proportional odds. Not all of the models supported the proportional odds assumption, which is a claim that the predictor variables have an equal impact on the outcome variable at each score level. However, this may not necessarily be problematic when estimating *average* odds ratios across samples/raters, and when the aim of the study is not to predict discrete outcomes for individuals. As Harrell (2020) wrote,

> When [the proportional odds assumption] does not hold, the odds ratio from the proportional odds model represents a kind of average odds ratio… a unified [proportional odds] model analysis is decidedly better than turning to inefficient and arbitrary analyses of dichotomized values of Y (Conclusion section, para. 1).

For this reason, I estimated model effects using mixed effects ordinal regression rather than comparable multinomial models, which would hinder parsimony and interpretability.

In each model, I entered variables by order of correlation with the dependent variable, creating four

models including the null model. Each model began with the null model, entered as

clmm(Language Score~1+(1|Rater)+(1|Sample))

where Language Score refers to each dependent variable (fluency, grammar, vocabulary, and comprehensibility), 1 indicates that no fixed effects were entered, and rater and sample were entered as separate random effects. Following this, the remaining variables were entered in each model one by one, with the final model including all three:

$clmm(Language\ Score \sim Var_1 + Var_2 + Var_3 + (1|Rater) + (1|Sample))$

All models used a logit link and flexible thresholds for the most accurate estimation of score probabilities. I then selected the best fitting model based on comparisons using likelihood ratio tests. I also tested the final selected model against the same model with random effects removed (a *clm* model) to verify that random effects contributed meaningfully to the model. Bonferroni corrections were applied for the four sets of analyses to control for multiple hypothesis tests, resulting in a more conservative Type I error threshold of $\alpha = .0125$. Plots of regression slopes, with points jittered to avoid overlap at scale scores, are presented to illustrate relationships with the final scores.

Interactions. Interactions between base language proficiency level (using rescaled IELTS scores) and the behavioral variables was also conducted in *clmm*. In this model *Prof* refers to the base proficiency, scaled IELTS score, and interactions were tested against all measured variables in the model simultaneously. The model is as follows:

$clmm(Language\ Score \sim Prof*Var_1 + Prof*Var_2 + Prof*Var_3 + (1|Rater) + (1|Sample))$

Significant interactions were explored using the *marginaleffects* package in R (version 0.8.1). However, the post hoc analysis using this software required two key modifications. This package computes marginal effects of interaction terms for cumulative logit models without mixed effects—*clm* models. I used *clm* models for post hoc analyses using this package because *clmm* lacks the *predict* function, making it impossible to run *marginaleffects* (Arel-Bundock, personal communication), and thus this method is the best approximation to analyze these interactions. The second modification is that *marginaleffects* can only handle categorical interaction terms. I thus dichotomized proficiency into two levels, low and high, using

the rescaled IELTS scores. I decided not to split proficiency into three levels (low, medium, and high), as this would have created two groups with only 8 cases, and 15 samples per case in the dichomotomized interaction was already a limited size. The low proficiency group included scaled scores below 4, leading to 14 samples equivalent to IELTS scores below 5.5, which is accepted to be below B2 level on the CEFR (Lim et al., 2013). High proficiency included scores greater than or equal to 4, which accounted for 16 scores at IELTS 5.5 or above, generally considered to be at B2 or above. The interaction models were then:

*clm(Language Score~Prof*Interaction_Var)*

I then used the *comparisons* function to compute the coefficients for each comparison. This function gives comparisons for proficiency level at each score level, 1–7. This gives insight into the differential effects of the interaction according to the scores raters assigned each sample. I used the *plot_cap* function to visualize the conditional adjusted predictions of the interactions, which lists the predicted probabilities of a score assignment for a given proficiency level for a single interaction variable.

Stimulated recall. After transcribing the stimulated recall data, I devised a coding scheme to begin analyzing the dataset. I first began by reading a sample of transcript files carefully and assigning codes based on items from the rating scales, behaviors identified in the ELAN file preparation, and other topics the raters identified (comments on content of what was discussed, desire to communicate, humor, etc). This resulted in a set of 40 preliminary codes. In the process I describe below, these codes were reduced to 38 and arranged thematically before being applied to the whole dataset.

I began by segmenting the dataset into idea units based on a semantic analysis of the textual content. I used Brown et al.'s (2005) definition of idea units as "single or several utterances, either continuous or separated by other talk but falling within the same turn, with a single aspect of performance as the focus" (p. 14). Most recalls consisted of a single idea unit, though some longer recalls consisted of multiple. I segmented rater 9's recall in the following example, where there was a comment on sample 9 after playing the file for 1 minute. This recall was complex and contained multiple ideas within. Idea units in this example are indicated with double slashes (//), which mark idea unit boundaries. Three idea units appeared in this recall. Note that these idea units contain multiple references to concepts, but the idea itself is generally

focused.

> // **(Unit 1)** *Um, I would say I know that I put vocabulary pretty strong. And I think it was not necessarily just from the beginning, but throughout the video, I could tell that she was saying some words that I may have not expected someone who doesn't know English very well to use, which is why I also put fluency higher, because from that, I would assume that she knows English better than other people.* // **(Unit 2)** *And I was also comparing these videos to each other a lot, which I think probably impacted what my scores were like, from the first videos I watched towards the ones at the end, because I was able to compare them to other ones. I don't, I'm not sure when I watched this video in comparison to the others.* // **(Unit 3)** *But I don't know from what I can tell so far, even from just like the first sentence of what she said, it seems like she knew what she was saying. And she understood the question very well.* //

The first idea unit focused on the relationship between overall language ability, vocabulary, and fluency. The second idea unit was an observation on metacognitive rating strategies, namely comparing performances and the effects of order on the rater's scores. The final idea unit contained observations about the content of the utterance and its relationship to test question comprehension.

I segmented the entire dataset following the above scheme alone because segmentation "require[s] subjective interpretation, contextualization, and especially a thorough understanding of the theoretically motivation questions guiding the study" (Campbell, et al., 2023, p. 304). Following segmentation, I began refining the coding scheme I had developed earlier. I first ran a reliability check to determine whether my codes were logical and applicable to this dataset. For the check, I input 10% of the idea units—a total of 186 units—into an Excel sheet for double coding. This percentage was chosen rather than the more standard 20% due to the substantial size of the dataset (with precedent in the literature, e.g., Sato & McNamara, 2019). The exact agreement between the raters was 77%. Based on the areas of disagreement, I refined the coding scheme to reduce ambiguity; for example, initially I had included codes for various aspects of content (amount of discussion, breadth of discussion, naturalness of content, truthfulness of ideas), which I collapsed to one *test content* category. Afterwards, the research assistant and I discussed the remaining

areas of disagreement, and each made final decisions on the remaining idea units. The final agreement rate was 98%. The final coding scheme I developed is presented in Table 4.5.

Table 4.5
NVivo Coding Scheme

Category	Code	Category	Code
Affect	• *Anxiety* • *Attentiveness* • *Attitude* • *Competence* • *Confidence* • *Desire to Communicate* • *Engagement* • *Expressiveness* • *Happiness* • *Humor* • *Interactiveness* • *Warmth*	Behavior	• *Eyebrows* • *General Face Behaviors* • *General Body Language* • *Gaze* • *Gesture* • *Head* • *Mouth* • *Paralinguistics* • *Posture*
Test Interaction	• *Active listening* • *Content* • *Examiner* • *Relevance-Contingence* • *Repair* • *Thinking* • *Turn-taking* • *Visual Artifacts*	Language	• *Comprehensibility* • *Comprehension* • *Fluency* • *Grammar* • *Organization* • *Overall ability* • *Pronunciation* • *Vocabulary*

I coded the entire dataset in NVivo for Mac (Version 12). First, I assigned attribute coding, identifying sections of text by stimulated recall rater and by test taker. This resulted in 200 unique observations (20 raters by 10 stimulated recalls). Following this, I coded the entire dataset using the coding scheme in Table 4.5. This resulted in 4,251 decisions on 1,213 idea units (M = 3.5 codes per idea unit). To illustrate, I provide an example of the coding process in Table 4.6. Unit 1 only contained language-related evidence, so it was coded as vocabulary, fluency, and overall ability. Unit 2 did not contain any information of relevance to this study, as mentions of metacognitive strategies were beyond the scope of this analysis. Thus, Unit 2 was not coded. Unit 3 contained multiple codes, namely a focus on comprehension (language), the content of the utterance (they knew *what* they were saying; test interaction), and an implication that the test taker was competent (they *knew* what they were saying; affect).

Table 4.6
Coding Example

Unit	Attribute		Code
1	Rater 9 / Sample 9	*Um, I would say I know that I put vocabulary pretty strong. And I think it was not necessarily just from the beginning, but throughout the video, I could tell that she was saying some words that I may have not expected someone who doesn't know English very well to use, which is why I also put fluency higher, because from that, I would assume that she knows English better than other people.*	**LANGUAGE** (Vocabulary, fluency, overall ability)
2	Rater 9 / Sample 9	*And I was also comparing these videos to each other a lot, which I think probably impacted what my scores were like, from the first videos I watched towards the ones at the end, because I was able to compare them to other ones. I don't, I'm not sure when I watched this video in comparison to the others*	N/A
3	Rater 9 / Sample 9	*But I don't know from what I can tell so far, even from just like the first sentence of what she said, it seems like she knew what she was saying. And she understood the question very well.*	**TEST INTERACTION** (Content) **LANGUAGE** (Comprehension) **AFFECT** (Competence)

While these codes allowed for a broad overview of where raters directed their attention in the stimulated recalls, sufficient for analyzing patterns underlying the relationship between nonverbal behavior and language proficiency (RQ3.2), an extra level of granularity was necessary to identify the nonverbal behaviors raters found most salient during their recalls (RQ3.1). Once the entire dataset was coded, I added an additional layer of subcodes to the Behavior category based on the raters' comments. These nonverbal behavioral subcodes are listed in Table 4.7, and resulted in 505 additional coding decisions.

Table 4.7
Subcodes of Nonverbal Behavior

Code	Subcode	Code	Subcode
Eyebrows	• Furrowed • Movement • Raised	Gaze	• Averted • Blinking • Eyes grow wide • Mutual • Shifting (Movement) • Staring • Unfocused
General Face Behaviors	• *No subcodes*	Gesture	• Lack of hand movement • Random movement • Representational gestures
General Body Language	• *No subcodes*		• Self-adaptors
Head	• Turns • Nodding	Mouth	• Frowning • Lack of smile • Lip movements • Mouth barely open • Nervous smile • (Genuine) smile • Swallowing
Paralinguistics	• Audible breathing • Backchanneling • Filled pauses • Laughing • Speed • Tone-prosody • Volume	Posture	• Adjusting posture • Leaning back/Slouching • Leaning forward • Moving around • Rigid/Straight posture • Rocking-Shaking • Shoulder movements

The coding scheme allowed me to view the frequencies of appearance of different comments regarding nonverbal behavior and affect, the extensiveness of their appearance across the 20 raters, and the relationships between comments and judgements of language. By analyzing areas where language ratings intersected with behavior, I was able to extract patterns and themes from the dataset. I generally followed Corbin and Strauss' (2015) method of constant comparisons, whereby intersections of data are repeatedly checked for similarities and differences in order to build theory. This analysis, however, deviates from Corbin and Strauss' grounded theory in that I approached the topic with clear hypotheses. I present the

themes, frequencies, and extracts illustrating the patterns I found in Chapter 7. In order to illustrate further the observations of the raters, where appropriate, I also include the multimodal transcripts from ELAN alongside quotes.

CHAPTER 5: AFFECT AND LANGUAGE PROFICIENCY

The purpose of this chapter is to describe methods used to analyze rating data from the online survey. First, I will describe the method I used to enhance data integrity by removing participants that exhibited scoring tendencies that were irregular or undesirable. Second, I will describe the dataset of observed rater judgements, including scores on language elements and impressionistic judgements, with the aim of providing evidence that the online survey provided meaningful data. I will then consider the structure of the dataset through an exploratory factor analysis to determine whether the 14 rating categories showed an underlying factor structure that could be used to reduce the dimensionality of the dataset. Finally, I will use the reduced dataset to explore relationships amongst the observed variables. The research question guiding Chapter 5 is:

RQ1: What is the relationship between interpersonal affect and language proficiency?

Participant selection

In the planning of this study, it was anticipated that some raters would exhibit undesirable rating tendencies. More raters were recruited than necessary so that problematic participants could be removed, but it was also desirable to keep as many raters as possible and to adopt fairly liberal measures to do so. It was expected that raters would exhibit irregularities given that they were not trained and only had minimal practice before completing the study. This allowed the integrity of the dataset to be strengthened without losing substantial statistical power. I considered three key threats to rating quality: 1) raters with low correlations against other raters on each language criterion, 2) rater misfit, and 3) multivariate outliers. The below analyses led to the exclusion of 16 raters.

Spearman correlations

The mean Spearman correlations resulted in 63 raters that correlated in their ranking of the test takers at .5 or above, and 89 raters that correlated at .4 or above. Given the exploratory nature of this study, I opted for the less conservative correlation estimate in order to retain the higher number of raters. Table 5.1 shows the correlations of the seventeen raters with the lowest grand means, of which 11 had means

below .4. These 11 raters were flagged to be removed from the dataset.

Table 5.1
Lowest Means and Grand Means of Spearman Correlations

Rater	Fluency	Vocabulary	Grammar	Pronunciation	Grand Mean
86*	.24	.17	.21	.20	.20
72*	.15	.30	.29	.14	.22
16*	.36	.33	.15	.15	.25
14*	.46	.32	.37	.28	.36
50*	.45	.45	.20	.34	.36
94*	.39	.45	.18	.47	.37
47*	.46	.35	.33	.37	.38
37*	.43	.35	.31	.44	.38
96*	.50	.42	.22	.42	.39
21*	.50	.41	.34	.34	.39
58*	.53	.45	.21	.35	.39
44	.49	.44	.39	.31	.41
30	.42	.41	.39	.44	.41
78	.47	.46	.27	.43	.41
51	.48	.44	.25	.46	.41
68	.51	.42	.40	.29	.41
15	.54	.45	.36	.34	.42
53	.63	.55	.02	.49	.42

Note. * indicates rater's grand mean was less than .4 and was dropped from the study.

Misfit

The second criterion for exclusion was the presence of possible rater effects visible through misfit in a MFRM model. Figure 5.1 shows the fit measures of the upper and lower end of the rater measurement table. It can be seen that 8 raters underfit the model, with infit over 1.5, and only one rater overfit the model with fit indices below .5. Although infit values lower than .5 indicate more stable than expected rating patterns, there was only one rater in this category, so this individual was left in the dataset. Three of these raters were also in the set of raters with low correlations. These five additional raters were flagged to be removed from the dataset.

Figure 5.1
Fit Measures of Raters

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
589	120	4.91	5.11	-.71	.08	2.00	6.0	2.39	7.6	-.04	.28	.67	21.9	25.9	72	72
505	120	4.21	4.25	-.17	.08	1.91	5.8	2.10	6.5	.11	.43	.70	22.4	25.7	94	94
544	120	4.53	4.67	-.41	.08	1.88	5.6	1.88	5.4	.28	.31	.69	20.6	26.3	86	86
541	120	4.51	4.64	-.39	.08	1.63	4.2	1.70	4.4	.56	.57	.69	24.1	26.3	30	30
579	120	4.82	5.02	-.64	.08	1.64	4.2	1.51	3.3	.57	.69	.68	23.7	26.1	46	46
463	120	3.86	3.79	.09	.08	1.41	3.0	1.62	4.1	.32	.50	.70	20.9	24.3	14	14
477	120	3.97	3.95	.00	.08	1.46	3.3	1.62	4.1	.58	.58	.70	24.0	24.8	66	66
505	120	4.21	4.25	-.17	.08	1.55	3.8	1.61	4.0	.42	.72	.70	23.3	25.7	82	82
495	120	4.13	4.14	-.11	.08	1.52	3.6	1.57	3.8	.63	.60	.70	24.6	25.4	71	71
561	120	4.68	4.84	-.52	.08	1.54	3.6	1.46	3.1	.59	.71	.69	25.1	26.3	80	80
528	120	4.40	4.50	-.31	.08	1.42	3.0	1.48	3.2	.74	.57	.69	25.7	26.1	68	68
606	120	5.05	5.26	-.83	.08	1.27	1.9	1.46	3.0	.54	.57	.66	24.1	25.5	21	21
569	120	4.74	4.92	-.57	.08	1.39	2.7	1.45	3.0	.77	.56	.68	25.6	26.2	10	10
587	120	4.89	5.09	-.69	.08	1.43	2.9	1.35	2.4	.67	.75	.67	24.3	25.9	56	56
660	120	5.50	5.70	-1.23	.09	1.42	2.7	1.32	2.1	.79	.57	.62	23.7	23.1	51	51
594	120	4.95	5.16	-.74	.08	1.23	1.7	1.36	2.4	.76	.62	.67	24.3	25.8	84	84
532	120	4.43	4.54	-.34	.08	1.26	1.9	1.36	2.5	.98	.66	.69	28.4	26.2	55	55
519	120	4.32	4.40	-.26	.08	1.17	1.3	1.33	2.3	.66	.58	.70	24.7	26.0	96	96
.																
.																
.																
529	120	4.41	4.51	-.32	.08	.67	-2.9	.65	-3.0	1.58	.71	.69	30.4	26.2	12	12
599	120	4.99	5.20	-.78	.08	.65	-3.0	.64	-3.0	1.36	.73	.67	28.8	25.7	61	61
661	120	5.51	5.71	-1.24	.09	.64	-2.9	.64	-2.9	1.30	.71	.62	24.7	23.0	88	88
512	120	4.27	4.33	-.21	.08	.63	-3.3	.73	-2.2	1.09	.70	.70	26.2	25.9	8	8
445	120	3.71	3.60	.20	.08	.63	-3.4	.75	-2.1	1.12	.67	.69	23.4	23.4	6	6
530	120	4.42	4.52	-.32	.08	.62	-3.4	.64	-3.1	1.31	.75	.69	28.8	26.2	41	41
546	120	4.55	4.69	-.43	.08	.59	-3.7	.58	-3.7	1.43	.76	.69	29.4	26.3	11	11
558	120	4.65	4.81	-.50	.08	.59	-3.7	.58	-3.6	1.45	.84	.69	29.9	26.3	3	3
540	120	4.50	4.63	-.39	.08	.56	-4.1	.65	-3.0	1.11	.68	.69	26.8	26.3	99	99
561	120	4.68	4.84	-.52	.08	.57	-3.9	.55	-4.0	1.47	.78	.69	29.1	26.3	62	62
479	120	3.99	3.97	-.01	.08	.60	-3.8	.54	-4.2	1.54	.79	.70	28.4	24.9	73	73
558	120	4.65	4.81	-.50	.08	.57	-3.9	.53	-4.3	1.48	.80	.69	29.2	26.3	48	48
649	120	5.41	5.62	-1.14	.09	.63	-3.1	.52	-4.1	1.48	.70	.63	27.1	23.7	40	40
567	120	4.72	4.90	-.56	.08	.50	-4.7	.55	-4.0	1.40	.71	.68	28.9	26.2	7	7
511	120	4.26	4.32	-.21	.08	.50	-4.9	.51	-4.5	1.35	.79	.70	27.8	25.8	28	28
595	120	4.96	5.16	-.75	.08	.42	-5.7	.45	-5.1	1.37	.76	.67	27.9	25.8	70	70
535.2	120.0	4.46	4.55	-.37	.08	1.00	-.2	1.04	.1		.68				Mean (Count: 99)	
56.9	.0	.47	.57	.38	.00	.32	2.5	.35	2.5		.10				S.D. (Population)	
57.2	.0	.48	.57	.38	.00	.32	2.5	.35	2.5		.10				S.D. (Sample)	

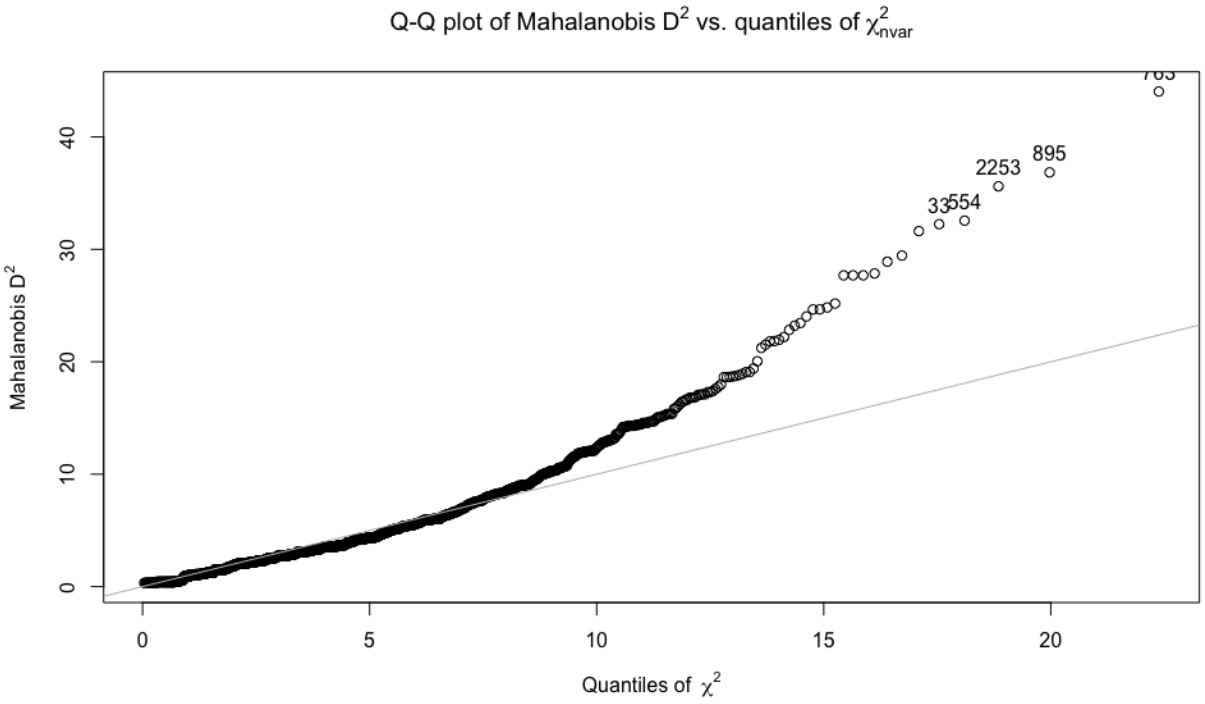
Model, Populn: RMSE .08 Adj (True) S.D. .37 Separation 4.55 Strata 6.40 Reliability (not inter-rater) .95
Model, Sample: RMSE .08 Adj (True) S.D. .37 Separation 4.57 Strata 6.43 Reliability (not inter-rater) .95
Model, Fixed (all same) chi-squared: 1954.3 d.f.: 98 significance (probability): .00
Model, Random (normal) chi-squared: 93.2 d.f.: 97 significance (probability): .59
Inter-Rater agreement opportunities: 582120 Exact agreements: 150454 = 25.8% Expected: 147701.4 = 25.4%

Multivariate outliers

The Q-Q plot of the Mahalanobis distances is shown in Figure 5.2. The numbers in this figure correspond to a single set of language scores (all four criteria) for one individual by one rater in the long form dataset. Points veering substantially from the diagonal line represent multivariate outliers. These were often cases of extremely jagged score profiles (such as 1, 1, 7, 1 for fluency, vocabulary, grammar, and comprehensibility), which could have been due to carelessness not noticing the shifting of polarity of the

semantic differentials. Each rater produced 30 ratings, so there were 2970 score profiles in total.

Figure 5.2
Q-Q Plot of Values and Multivariate Outliers



I then compiled a list of the raters to whom the multivariate outliers were associated. The frequency table of raters and the number of outliers associated with them is presented in Table 5.2. Each rater awarded 120 language ability scores (30 samples x 4 criteria), and the percentage of outlying scores is presented in the third column. I speculated that a small degree of carelessness (one or two ratings) would probably not impact the overall dataset, and dropping raters over a small number of cases could be detrimental if the raters' ratings were otherwise careful. There were only two raters with more than two outliers (46 and 72), and both of these had been identified as misfitting in the Rasch analysis or correlation analysis. These raters produced 4 and 6 outliers respectively, and as such were flagged to be removed. The rest of the raters in Table 5.2 were kept in the study as they may have represented important aspects of the population being sampled (Tabachnik & Fidell, 2013).

Table 5.2
Frequency Table of Multivariate Outliers

Rater	Number of outliers	% of outlier ratings (of 30)
2	2	6.67%
9	1	3.33%
13	1	3.33%
14	1	3.33%
21	1	3.33%
25	1	3.33%
26	1	3.33%
29	1	3.33%
32	1	3.33%
46	4	13.33%
48	1	3.33%
55	1	3.33%
56	1	3.33%
63	2	6.67%
66	2	6.67%
68	1	3.33%
72	6	20%
80	1	3.33%
82	1	3.33%
83	1	3.33%
90	1	3.33%
92	2	6.67%
94	2	6.67%
97	1	3.33%

As mentioned previously, the cleaning analysis resulted in 16 raters being dropped from the dataset. Four of these raters showed low alignment with the greater group of raters through grand mean Spearman correlations (30, 71, 80, 82). Eight showed misfit in the Rasch model (14, 16, 21, 37, 47, 50, 58). Four additional raters showed a combination of either > 10% of scores being outliers, misfit, or low correlations (46, 86, 72, 94). Additionally, one rater had been dropped prior to the analysis due to technical problems in the recording. The remaining analyses were thus conducted with 83 raters.

Dataset description

Scales

Mean scores on each of the language and affect scales were situated towards the midpoint of four or slightly higher, with raters assigning the full range of scores for each scale. Descriptive statistics are presented in Table 5.3. Here we can see that the lowest mean score was grammar (4.14), while the highest

mean score was attention (5.34). Standard deviations indicated somewhat less variance for most of the affect-related scales, with the lowest variance in attitude (1.22), while language-related scores showed greater variance with a high *SD* in vocabulary (1.67). Alpha levels varied between .65 (anxiety) and .85 (fluency). These levels, although a little low for operational testing (especially anxiety), indicate a fair degree of consistency despite the unlabeled scale levels and lack of rater training. Indeed, .85 is remarkably good, and is on par with high-stakes, standardized tests (Nunnaly & Bernstein, 1994; Zhang, 2010). Interrater reliability estimates, estimated using intraclass correlation coefficients (ICC), will be presented in the section on raters below.

Table 5.3
Descriptive Statistics and Cronbach's Alpha of Scales

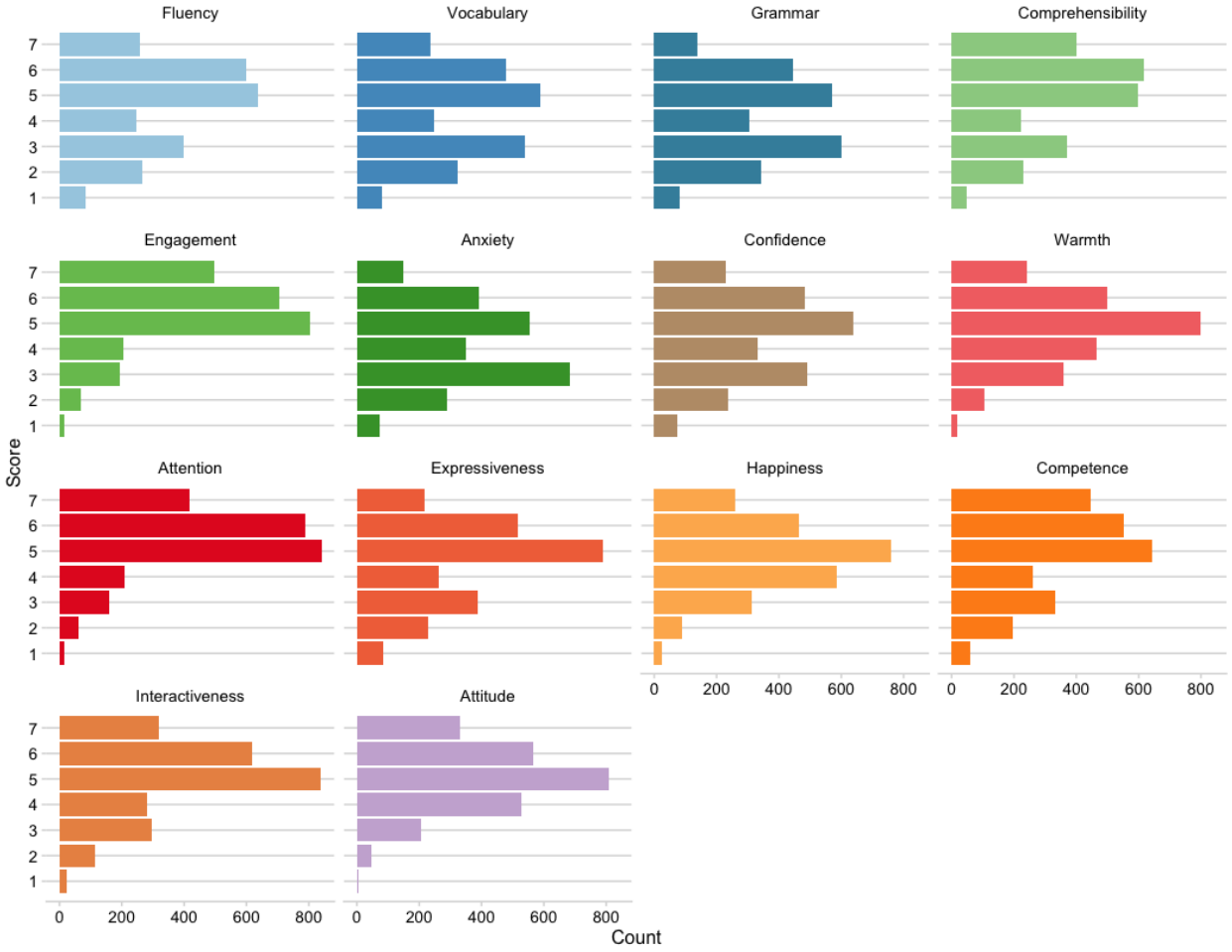
Scale	Mean	*SD*	Skewness	Kurtosis	*SE*	α
Fluency	4.58	1.65	-0.40	-0.86	0.03	.85
Vocabulary	4.33	1.67	-0.13	-1.07	0.03	.81
Grammar	4.14	1.59	-0.03	-1.03	0.03	.72
Comprehensibility	4.83	1.64	-0.47	-0.81	0.03	.79
Engagement	5.33	1.31	-0.79	0.37	0.03	.79
Anxiety	4.12	1.54	0.06	-0.92	0.03	.65
Confidence	4.45	1.59	-0.23	-0.86	0.03	.82
Warmth	4.76	1.34	-0.28	-0.43	0.03	.73
Attention	5.34	1.23	-0.83	0.71	0.02	.74
Expressiveness	4.56	1.57	-0.44	-0.64	0.03	.77
Happiness	4.77	1.32	-0.23	-0.30	0.03	.75
Competence	4.87	1.63	-0.51	-0.66	0.03	.84
Interactiveness	4.98	1.39	-0.58	-0.17	0.03	.78
Attitude	5.05	1.22	-0.23	-0.36	0.02	.74

Table 5.4 and Figure 5.3 show the distribution of rater choices of individual score categories. While raters used the whole range of possible scores, some scales such as engagement and attention showed negatively skewed and highly kurtotic patterns, while others such as warmth and happiness showed a distribution appearing more "normally" distributed (normal is in quotations as normality is not considered for ordinal data). For the four language categories, raters appeared more likely to avoid a midpoint of four and instead lean towards a more positive or negative end of the scale. In fact, a selection of 4 was never the most common choice on any of the scales, even though it was the default selection on the scales in Qualtrics.

Table 5.4
Frequency Counts of Scale Categories

Scale	1	2	3	4	5	6	7
1. Fluency	82	265	400	247	637	599	260
2. Vocabulary	82	322	541	246	587	477	235
3. Grammar	83	342	600	307	572	447	139
4. Comprehensibility	49	230	370	224	598	617	402
5. Engagement	17	70	194	204	803	706	496
6. Anxiety	74	289	682	349	555	392	149
7. Confidence	75	237	492	331	640	485	230
8. Warmth	18	106	360	466	799	501	240
9. Attention	14	61	161	209	840	787	418
10. Expressiveness	83	229	389	263	790	517	219
11. Happiness	23	89	311	584	761	463	259
12. Competence	61	196	334	259	644	551	445
13. Interactiveness	25	115	295	279	837	620	319
14. Attitude	4	45	206	529	810	567	329

Figure 5.3
Scale Histograms



Scale correlations

Next, I considered the relationship amongst the scales. All correlations were positive, ranging from medium (.4) to strong (≥ .6) (Plonsky & Oswald, 2014). The weakest correlation was between anxiety and attention (.40), while the strongest correlation was between fluency and vocabulary (.85). The full correlation table is detailed in Table 5.5.

Table 5.5
Polychoric Correlations

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Fluency													
2. Vocabulary	.85												
3. Grammar	.75	.74											
4. Comprehensibility	.81	.74	.67										
5. Engagement	.63	.58	.51	.59									
6. Anxiety	.56	.54	.48	.50	.43								
7. Confidence	.73	.70	.61	.63	.62	.72							
8. Warmth	.48	.45	.41	.50	.63	.43	.54						
9. Attention	.59	.55	.47	.57	.84	.40	.58	.59					
10. Expressiveness	.58	.55	.47	.58	.66	.47	.61	.72	.60				
11. Happiness	.52	.48	.42	.52	.62	.45	.58	.82	.59	.73			
12. Competence	.84	.77	.68	.76	.68	.54	.70	.53	.66	.59	.56		
13. Interactiveness	.63	.58	.50	.59	.76	.45	.62	.63	.72	.68	.63	.67	
14. Attitude	.51	.47	.42	.52	.69	.41	.58	.79	.63	.70	.82	.55	.64

Note. All correlations significant at $p < .05$.

Scale functioning in Rasch

The model used to interpret scale functioning was the same partial-credit model as in the outlier analysis with samples, language criteria, and raters as facets. I did not run the model with the affect variables included due to possible problems with multidimensionality (Aryadoust et al., 2021). Overall, the language-related scales appeared to function as anticipated in an MFRM analysis of the final, outlier-removed dataset. The MFRM summary statistics are available in Table 5.6. Average fit statistics were all very close to 1.00 with standard deviations within recommended cutoffs of .5 to 1.5. The separability index was high at 10, suggesting that the scale could reliability separate ability levels into 10 different categories.

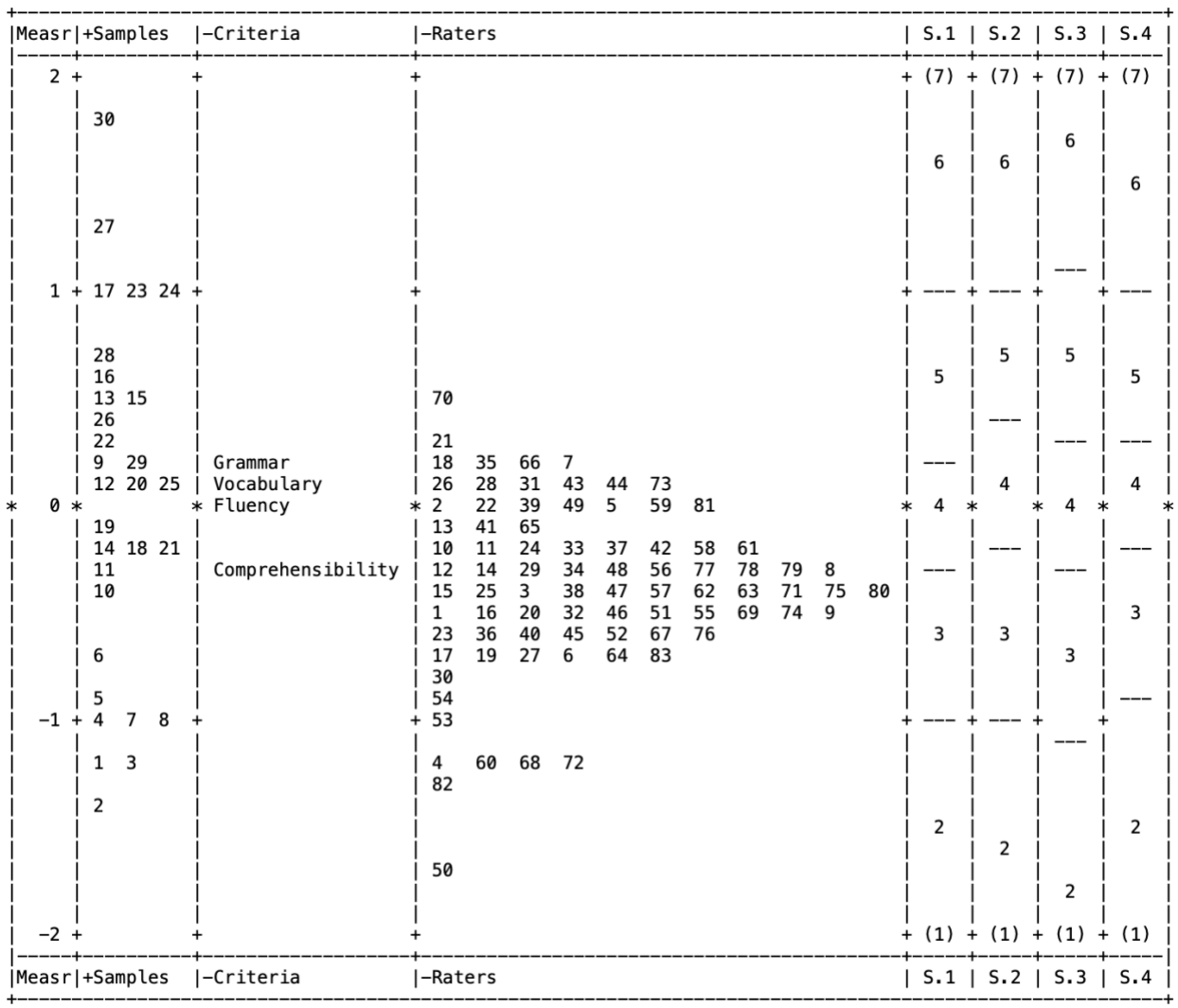
Table 5.6
MFRM Summary Statistics

	Rater	Sample	Criteria
N	83	30	4
Measures			
M	-0.38	0.00	0.00
SD (pop.)	0.59	1.12	0.35
SE	0.08	0.05	0.02
RMSE (pop.)	0.08	0.05	0.02
Adjusted (true) *SD* (pop.)	0.39	0.80	0.19
Infit *MS*			
M	1.00	1.03	1.01
SD (pop.)	0.31	0.20	0.13
Outfit *MS*			
M	1.04	1.04	1.04
SD (pop.)	0.33	0.21	0.16
Homogeneity index (χ^2)	1770.00	6363.40	458.40
df	82	29	3
p	< .001	< .001	< .001
Separation (pop.)	4.75	15.89	10.72
Reliability of separation (pop.)	.96	> .99	.99
Interrater reliability			
Observed exact agreement %	25.7		
Expected %	25.3		
Rasch κ	.001		

The Wright map for the final dataset on four language categories is presented in Figure 5.4. The figure shows that Grammar indeed was the more difficult criterion, as suggested by Table 5.3, and Pronunciation was the easiest. These differed by .52 logits. Each criterion demonstrated adequate measurement properties based on their category statistics. I have included these statistics and their category

probability curves in Appendix H.

Figure 5.4
Wright Map of Dataset



Note. S.1 = Fluency, S.2 = Vocabulary, S.3 = Grammar, S.4 = Comprehensibility.

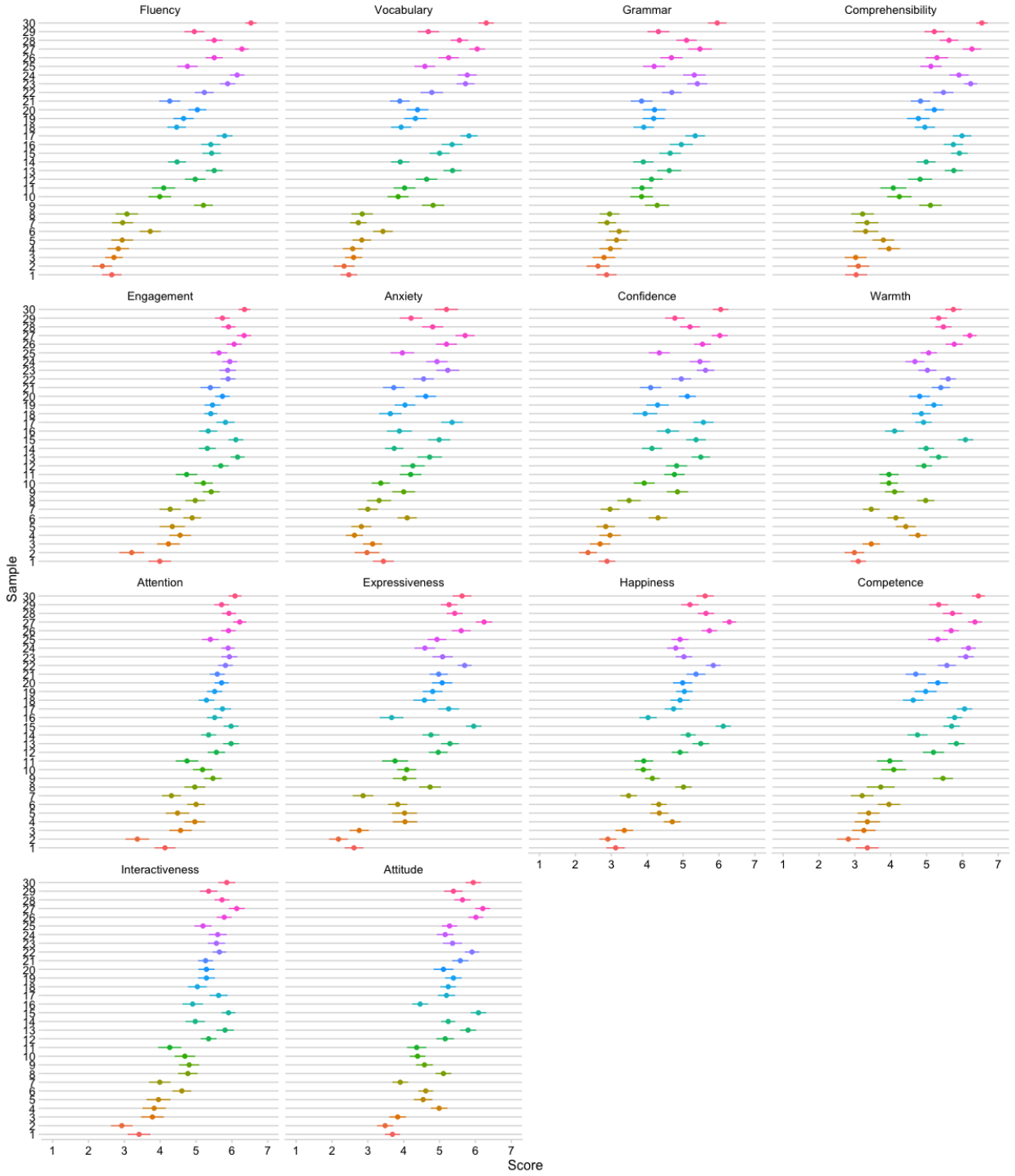
Samples

The means and confidence intervals for each sample by rating scale are displayed in Figure 5.5. Descriptive statistics of the ratings from the speech samples appeared to roughly correspond with their original ordering based on IELTS scores (Sample 1 was the least proficient candidate, while Sample 30 was the most), though as noted in the methods section, some samples dropped in their ranking substantially (e.g., Sample 29) while others rose (e.g., Sample 9). Surprisingly, this appeared to be the case for all rating

scales. Nevertheless, it can be seen that there was substantial variance across the criteria, with scales such as Engagement appearing much more linear than others such as Anxiety or Expressiveness. Similar to the scale MFRM statistics, average fit statistics for the samples were all very close to 1.00 with standard deviations within recommend cutoffs of .5 to 1.5, as seen in Table 5.6.

Figure 5.5

Means and CIs of Speech Samples



Raters

The final reduced set of raters exhibited rating patterns that fit the Rasch model well but showed substantial disagreement and variability. Average fitness parameters were close to 1, and standard deviations fell within recommended boundaries of .5 and 1.5. Fit statistics are dependent on the sample of individuals measured, however, so when the previously misfitting raters were removed, the truncated dataset showed some newly misfitting raters. Four additional raters misfit the model, as can be seen in Figure I.1 in Appendix I, but because this was anticipated, these additional raters were not removed. Regarding estimates of rater severity, The raters were as a group severe, with a mean logit score of -0.38. Figure I.2 shows that the vast majority of raters fell within a one logit range (-.50 – .50). There were, however, raters who were lenient to a large degree, up to -1.65 logits (rater 50). This resulted in a logit range of 2.17, and raters could be separated into 4.75 different severity levels overall.

Regarding consistency, raters showed disagreement. Exact rater agreement was 25.7%, only slightly higher than the Rasch expected agreement of 25.3%. This resulted in a Rasch Kappa index of .001, which is a model-expected level of agreement according to Linacre (n.d.). I also computed ICCs as a measure of interrater reliability. The ICCs, shown in Table 5.7, show that correlations for Fluency and Vocabulary were moderate, while the ICCs for Grammar and Comprehensibility were poor, especially for Grammar.

Table 5.7
Intraclass Correlation Coefficients

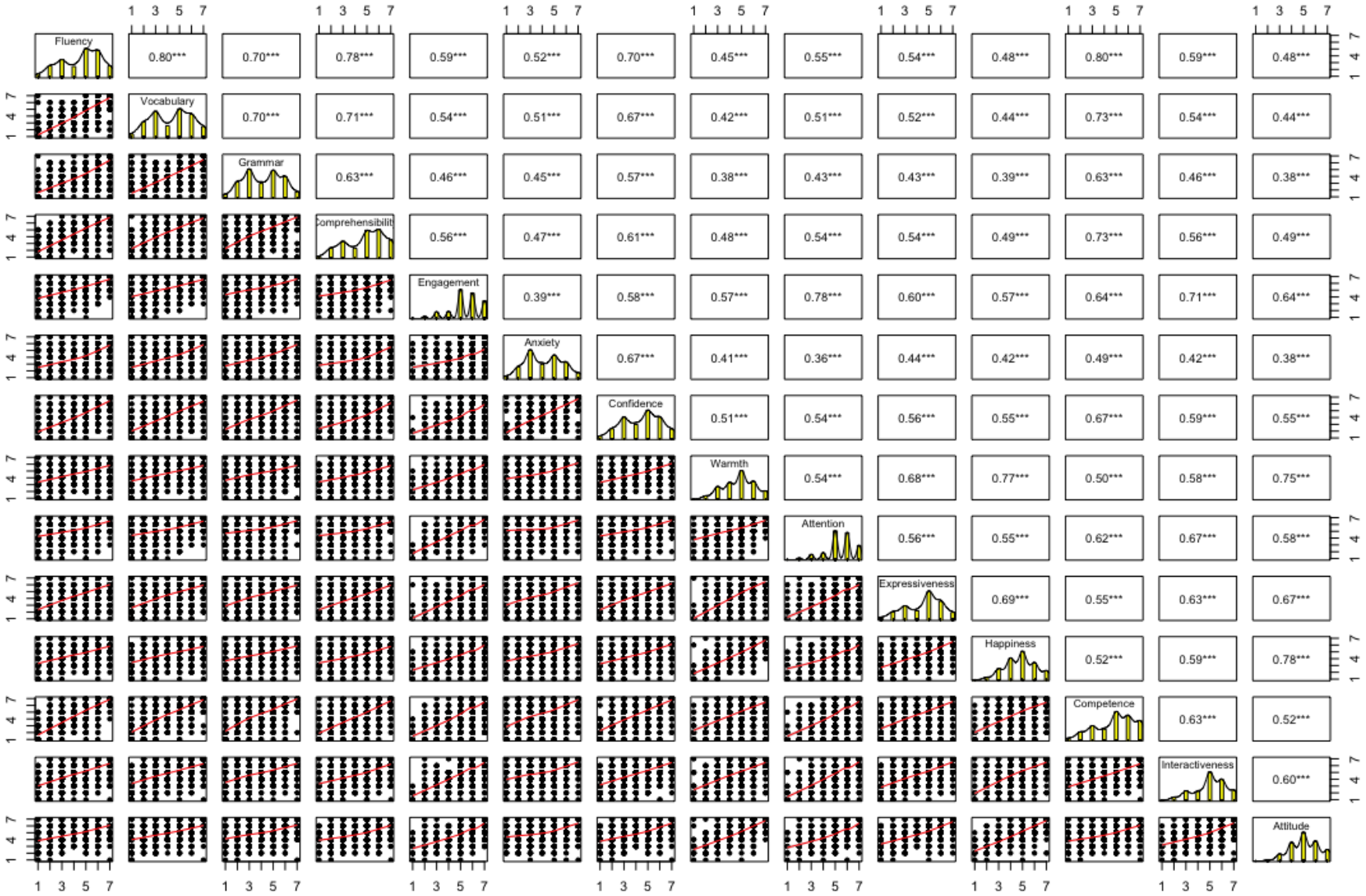
Outcome	ICC
Fluency	.58
Vocabulary	.54
Grammar	.37
Comprehensibility	.49

Data Structure

To understand the relationship amongst the variables measured in this study, I first ran an exploratory factor analysis to reduce the dimensionality of the scales. I first checked assumptions prior to conducting the factor analysis. The sample size of 83 with repeated measurements of 30 test takers resulted

in 2490 sets of observations, which is above the threshold suggested in Tabachnik and Fidell (2013). Correlations in Table 5.5 were above .30 but below .90, which suggests factorability and a lack of multicollinearity. I confirmed this by calculating the Kaiser measure of sampling adequacy, which was .95, and all scales exceeded .93. Furthermore, I checked for multicollinearity by building multiple linear regression models with all affect variables as predictors and the four proficiency variables as outcomes. All of the variance inflation factors (VIF) were lower than 4, and all VIF thresholds were higher than .25, which indicates the lack of collinearity issues. There was a linear relationship amongst all scales, as shown in Figure 5.6. Outliers were dealt with previously.

Figure 5.6
Linear Relationships Amongst Variables



A calculation of the eigenvalues resulted in only two values above 1, suggesting the presence of two factors using the Kaiser criterion. I used parallel analysis to further investigate this, as the Kaiser criterion is often a conservative estimate of the number of factors. Parallel analysis and the resulting Scree Plot in Figure 5.7 suggested four factors. Given that the dataset is ordinal, I used the polychoric correlation matrix for factor analysis rather than a Pearson correlation matrix. I then extracted four factors using exploratory factor analysis using maximum likelihood estimation and a promax rotation. The pattern matrix, shown in Table 5.8, indicates that communality values were quite strong, and all variables loaded on factors above common cutoff values of .45 with no substantial cross loading.

Figure 5.7
Scree Plot

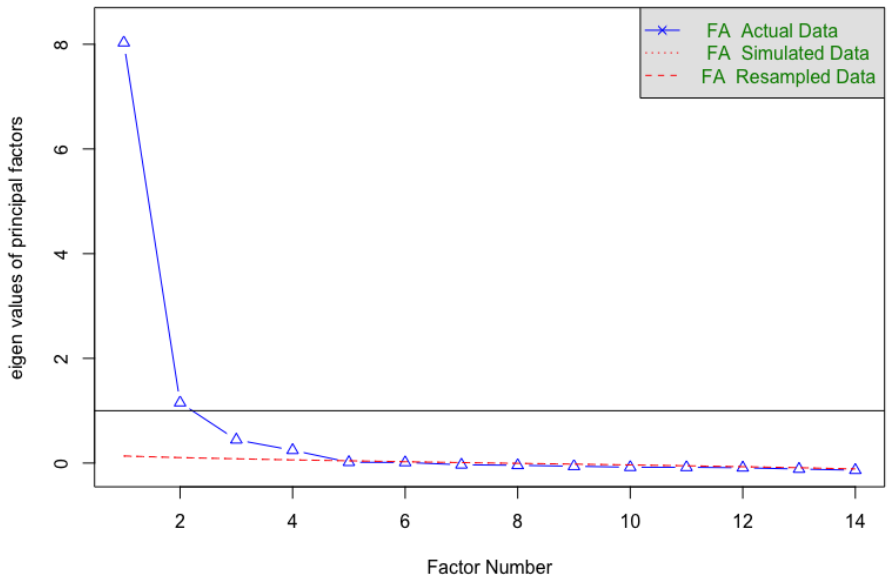


Table 5.8
Pattern Matrix

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Common	Unique
Fluency	**.96**	-.05	.01	.03	.91	.09
Vocabulary	**.88**	-.05	-.01	.06	.79	.21
Grammar	**.83**	-.02	-.06	.02	.63	.37
Comprehensibility	**.84**	.11	.00	-.08	.73	.27
Engagement	.00	.03	**.92**	.00	.88	.12
Anxiety	.17	.07	-.09	**.60**	.53	.47
Confidence	.02	-.04	.10	**.94**	.99	.01
Warmth	-.03	**.94**	-.02	-.03	.79	.21
Attention	.00	.02	**.88**	-.01	.79	.21
Expressiveness	.14	**.62**	.10	.03	.67	.33
Happiness	.01	**.99**	-.10	.00	.85	.15
Competence	**.74**	.01	.19	.00	.80	.20
Interactiveness	.14	.20	**.54**	.03	.69	.31
Attitude	-.08	**.82**	.13	.03	.79	.21
SumSq loadings	3.95	3.18	2.30	1.43		
Proportion variance	.28	.23	.16	.10		
Cumulative variance	.28	.51	.67	.78		

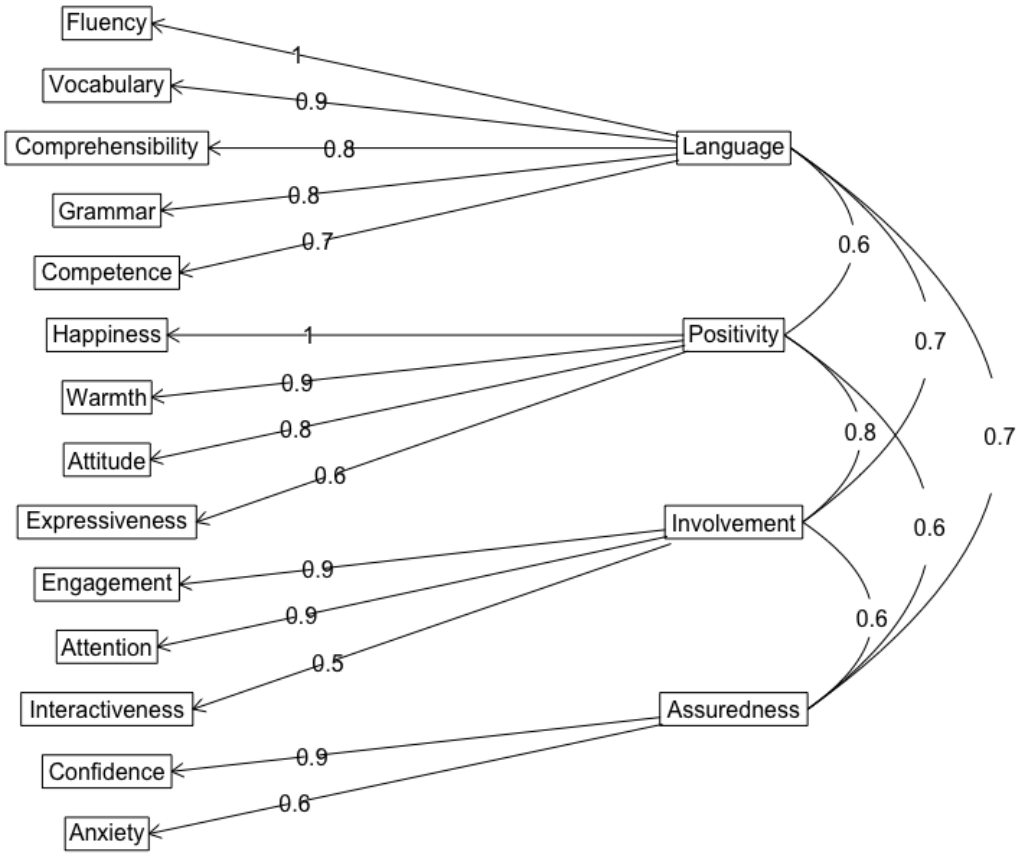
The factor loadings indicated a possible interpretation of the factor structure. Factor 1 consisted of strong loadings of fluency, vocabulary, grammar, comprehensibility, and competence. These factors together formed a factor I called *language*. The second factor consisted of strong loadings of warmth, happiness, and attitude, and a medium loading of expressiveness. I called this factor *positivity*, as it appeared to relate mostly to positive affect (warmth, happiness) indicated through behavior (expressiveness). The third factor was composed of strong relationships to engagement and attention, with a medium relationship to interactiveness. I called this factor *involvement*, as these three scales all related to the relationship the test taker established with the rater during the test. Finally, the fourth factor comprised a strong relationship with confidence and a medium relationship with anxiety. I called this factor *assuredness*, which approximates the relationship between these two affective states. Correlations amongst the four factors, shown in Table 5.9, were quite strong, suggesting either meaningful relationships between these categories or an artefact of rater effects (e.g., halo effect). The path diagram of the overall structure of the model is presented in Figure 5.8. Similar to path diagrams in confirmatory factor analysis, this model constrains the strongest factor loading to 1 to identify the factor, and for this reason the loadings and correlations look

slightly different from Table 5.8.

Table 5.9
Correlations Amongst Factors

	Language	Positivity	Involvement
Positivity	.63		
Involvement	.71	.76	
Assuredness	.74	.63	.60

Figure 5.8
Path Diagram of EFA Structure



Relationships between affect and proficiency measures

Because each affect-related factor correlated with language, I then investigated the relationship between each factor and the four language components of fluency, grammar, vocabulary, and comprehensibility. I did not investigate the relationship amongst these factors and competence as there was no theoretically informed reason to do so, so this variable was not included in the analysis. ICCs had

suggested that there was substantial variance amongst raters, and any regression model would require random effects accounted for at the second level (Tabachnik & Fidell, 2013). For this reason I used ordinal mixed-effects regression. For interpretability, the factor scores were scaled to span scores of 1–7 for comparability with the 7-point scales the raters used.

Fluency

Factors were entered in the model based on their correlations with fluency listed in Table 5.10. Assuredness had the strongest correlation with fluency, while positively had the lowest. The relationships are illustrated in Figure 5.9 using regression lines to best approximate the relationships across various levels of the predictors. It can be seen that while fluency increased relatively monotonically with assuredness, the relationships with the other factors were not as strong. The four models, shown in Table 5.11 indicated that the best fitting model was the one that only included assuredness, $\chi^2(1) = 74.58$, $p < .011$. This model fit significantly better than the model with random effects removed, $\chi^2(2) = 518.05$, $p < .001$, *clm* model AIC = 7458.10, *clmm* model AIC = 6944.10.

Table 5.10
Polychoric Correlations Between Factors and Fluency

Factor	Polychoric Correlation
Assuredness	.72 [69, 74]
Involvement	.70 [.67, .72]
Positivity	.57 [.55, .60]

Figure 5.9
Relationships Between Factors and Fluency

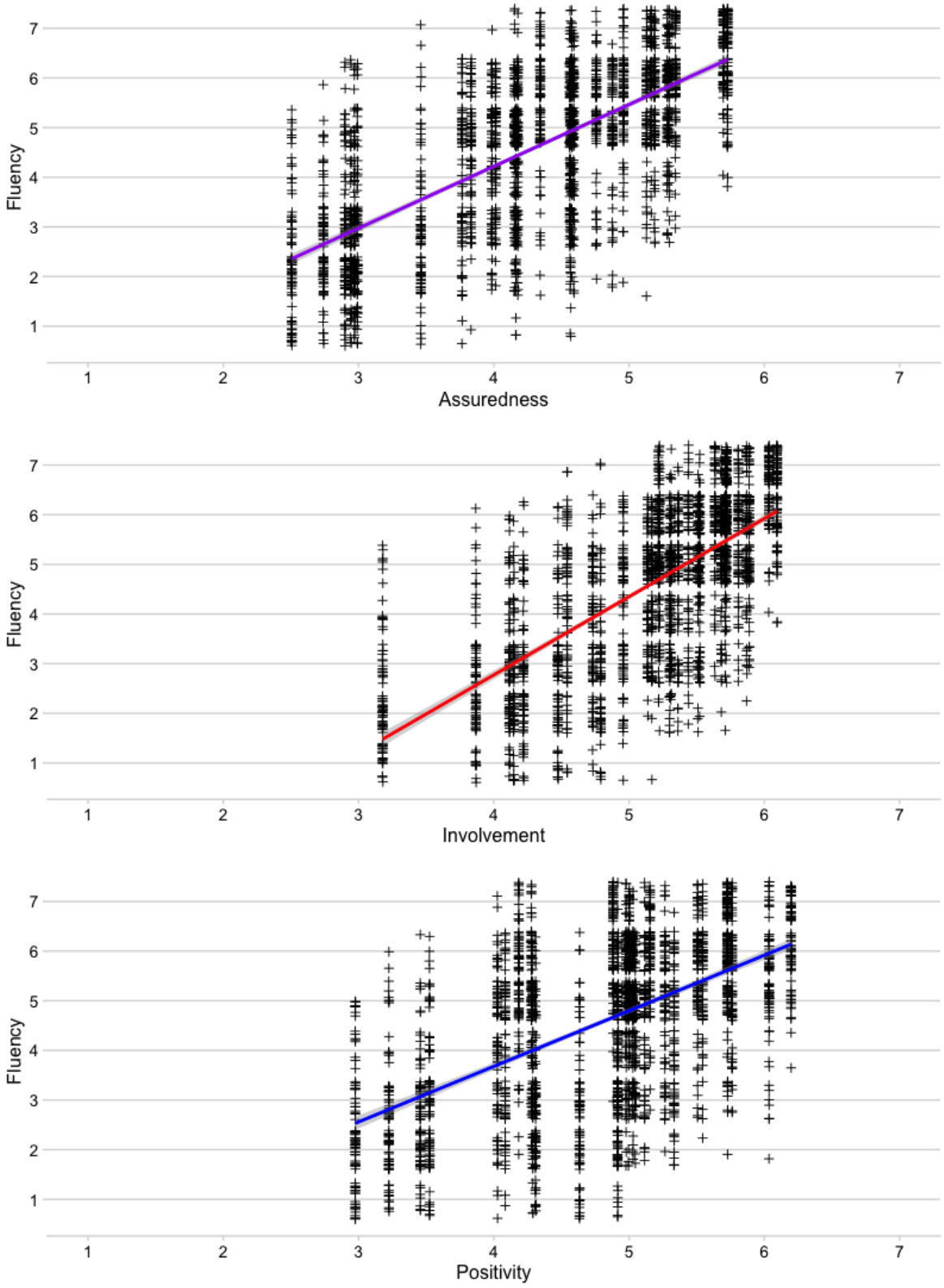


Table 5.11
Model Comparisons for Fluency

Model	AIC	χ^2	*df*	*p*
Null Model	7016.70			
Model 1	6944.10	74.58	1	< .001
Model 2	6942.40	3.65	1	.06
Model 3	6939.20	5.25	1	.02

Note. Adjusted α = .0125.

The final model, shown in Table 5.12, included only assuredness as a fixed effect. Assuredness significantly predicted learners' fluency, β = 3.09, p < .001. The odds ratio was 21.90, indicating that with each one-point change in assuredness, fluency was 21.90 times more likely to change in the same direction. There was more variance in raters, .97, than in samples, .32, which is not entirely unexpected, as raters were novice and untrained. Assuredness explained 20% of the variance in fluency scores, Nagelkerke's Pseudo R^2 = .20.

Table 5.12
Final Fluency Model

Coefficients	β	95% CI	*SE*	*z*	*p*	OR	95% CI
Assuredness	3.09	[2.74, 3.44]	.18	17.23	< .001	21.90	[15.42, 31.12]

Random effects

Groups	Variance	SD
Raters	0.97	0.98
Samples	0.32	0.57

Vocabulary

Factors were entered in the model based on the correlations with vocabulary listed in Table 5.13, which were similar and identical in order to the correlations with fluency. These relationships are illustrated in Figure 5.10. The regression graphs were very similar to those with fluency as well. The four models, shown in Table 5.14 indicated that the best fitting model was the one with all three factors, $\chi^2(1) = 7.61$, p = .006. This model fit significantly better than the model with random effects removed, $\chi^2(4) = 396.68$, p < .001, *clm* model AIC = 7638.80, *clmm* model AIC = 7250.20.

Table 5.13

Polychoric Correlations Between Factors and Vocabulary

Factor	Polychoric Correlation
Assuredness	.69 [.67, .72]
Involvement	.66 [.63, .68]
Positivity	.53 [.50, .56]

Figure 5.10
Relationships Between Factors and Vocabulary

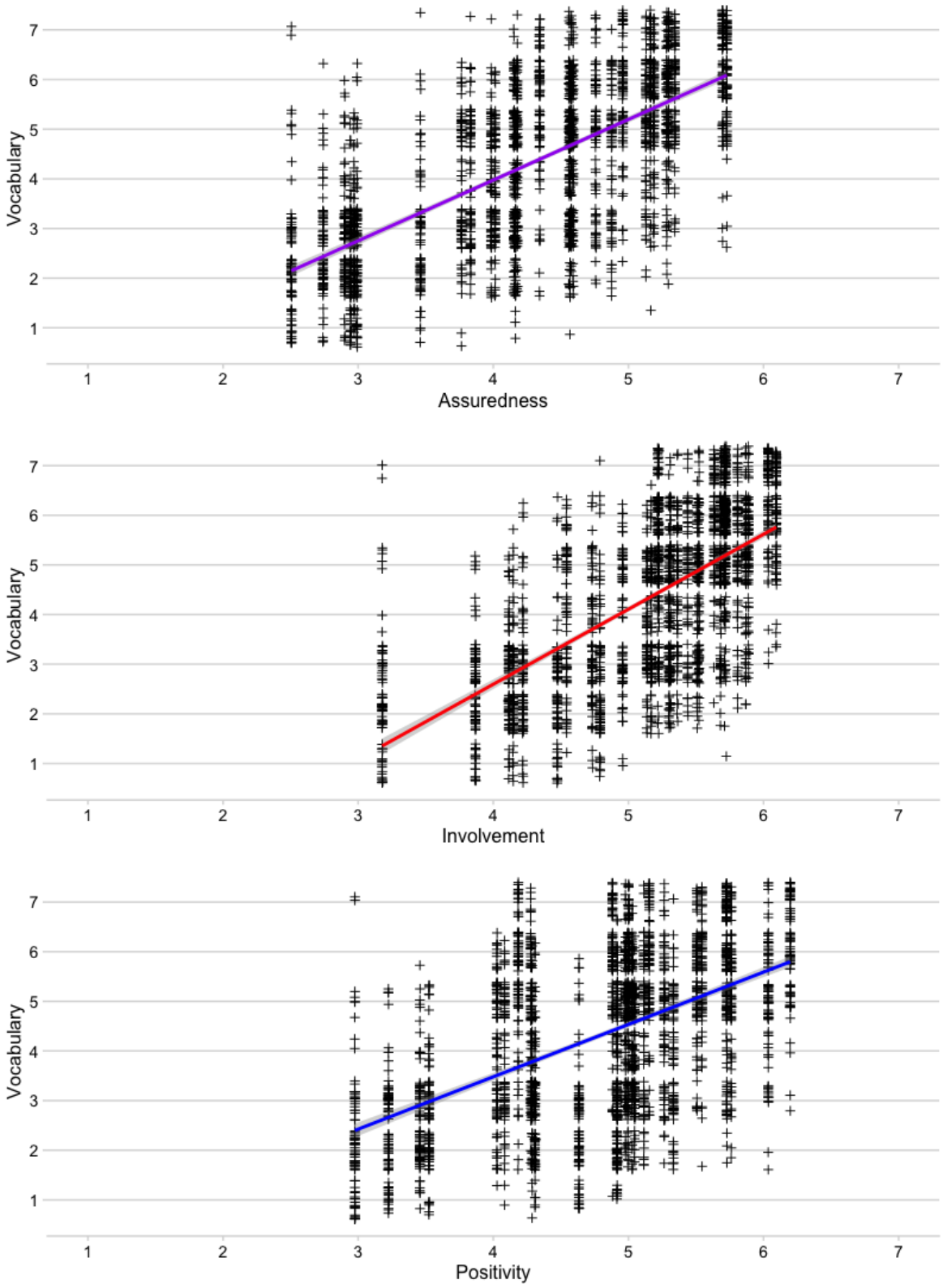


Table 5.14
Model Comparisons for Vocabulary

Model	AIC	χ^2	*df*	*p*
Null Model	7325.30			
Model 1	7255.50	71.84	1	< .001
Model 2	7255.80	1.71	1	.19
Model 3	7250.20	7.61	1	.006

Note. Adjusted α = .0125.

The final model, shown in Table 5.15, showed that all three factors predicted vocabulary. Unit changes in assuredness and involvement significantly predicted vocabulary scores (assuredness, β = 1.98, p < .001; involvement β = 1.90, p = .001), while changes in positivity had an inverse relationship with vocabulary scores (β = -0.96, p = .003). Odds ratios for assuredness and involvement indicated that one unit of change in each predictor resulted in between 6.67 and 7.27 times the likelihood of a change in vocabulary. The odds ratio for positivity was quite low at 0.38. There was less variance in raters in this model than the fluency model, 0.65, and likewise in samples, 0.21. The three factors explained 15% of the variance in vocabulary scores, Nagelkerke's Pseudo R^2 = .15.

Table 5.15
Final Vocabulary Model

Coefficients	β	95% CI	*SE*	*z*	*p*	OR	95% CI
Assuredness	1.98	[1.24, 2.73]	0.38	5.22	< .001	7.27	[3.45, 15.31]
Involvement	1.90	[0.74, 3.05]	0.59	3.23	.001	6.67	[2.11, 21.12]
Positivity	-0.96	[-1.60, -0.32]	0.33	-2.94	.003	0.38	[0.20, 0.72]

Random effects

Groups	Variance	SD
Raters	0.65	0.80
Samples	0.21	0.45

Grammar

Factors were entered in the model based on the correlations with grammar listed Table 5.16, which were ordered the same as fluency and vocabulary, but somewhat weaker. These relationships are illustrated in Figure 5.11. The relationships were also similar to fluency and vocabulary, especially for assuredness, but the association with involvement and positivity was weaker. The four models, shown in Table 5.17,

indicated that similar to fluency, the best fitting model was the one with only assuredness as a predictor, $\chi^2(1) = 59.61$, $p < .001$. This model fit significantly better than the model with random effects removed, $\chi^2(2) = 426.67$, $p < .001$, *clm* model AIC = 8130.30, *clmm* model AIC = 7707.60.

Table 5.16
Polychoric Correlations Between Factors and Grammar

Factor	Polychoric Correlation
Assuredness	.55 [.52, .58]
Involvement	.52 [.49, .56]
Positivity	.43 [.39, .46]

Figure 5.11
Relationships Between Factors and Grammar

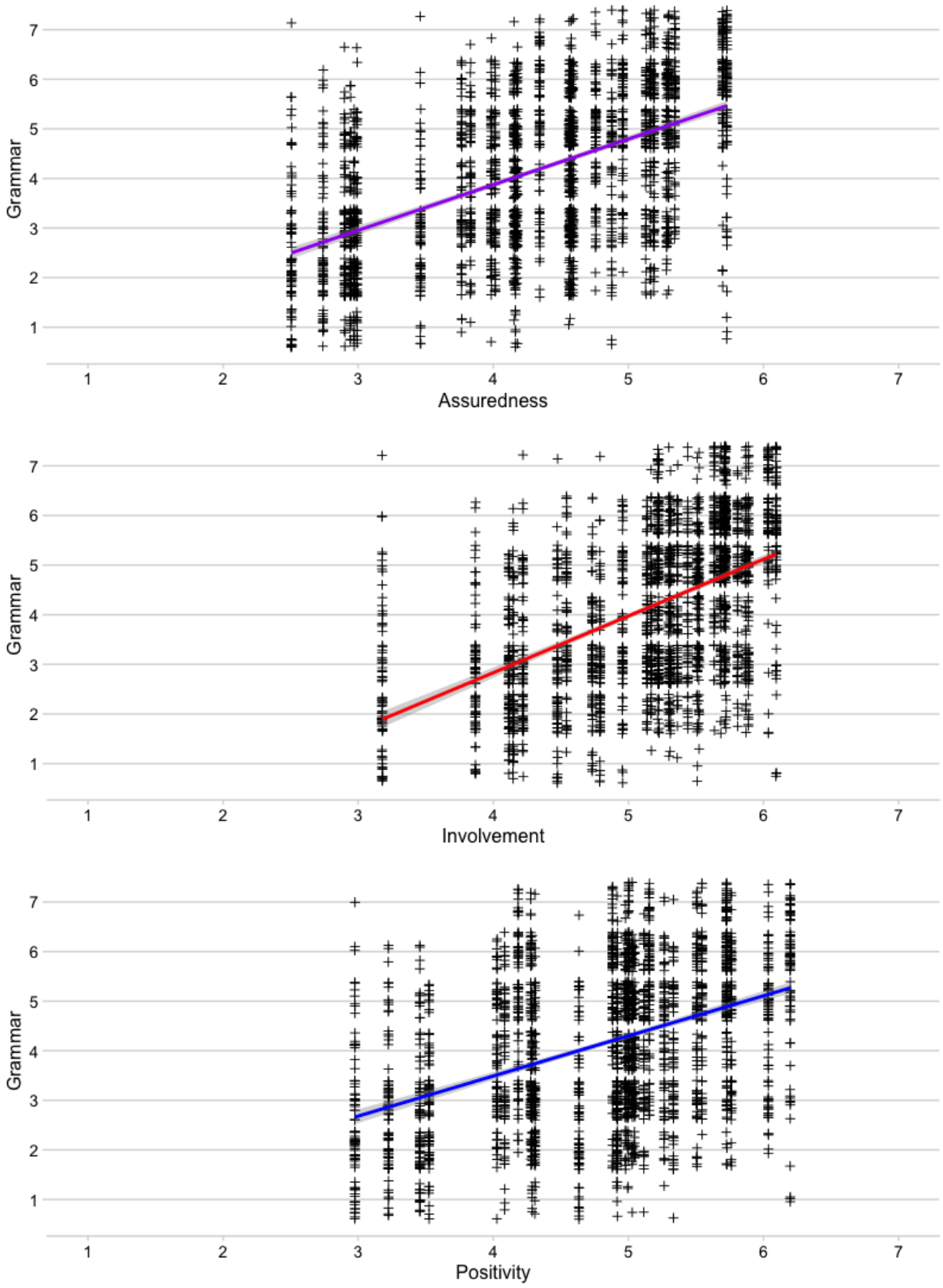


Table 5.17
Model Comparisons for Grammar

Model	AIC	χ^2	*df*	*p*
Null Model	7765.20			
Model 1	7707.60	59.51	1	< .001
Model 2	7708.30	1.31	1	.25
Model 3	7707.80	2.56	1	.11

Note. Adjusted α = .0125.

The final model, shown in Table 5.18, showed that only assuredness predicted grammar scores, β = 2.04, p < .001, indicating that with each one-point change in assuredness, grammar was 7.67 times more likely to change in the same direction. There variance in this model was similar to previous model, raters = 0.80, samples = 0.24. Assuredness explained 16% of the variance in grammar scores, Nagelkerke's Pseudo R^2 = .16.

Table 5.18
Final Grammar Model

Coefficients	β	95% CI	*SE*	*z*	*p*	OR	95% CI
Assuredness	2.04	[1.74, 2.34]	0.15	13.32	< .001	7.67	[5.69, 10.36]
	Random effects						
	Groups	Variance	SD				
	Raters	0.80	0.89				
	Samples	0.24	0.49				

Comprehensibility

Factors were entered in the model based on the correlations with comprehensibility listed in Table 5.19, which were ordered differently from previous models, with involvement coming first, followed by assuredness. The relationship between comprehensibility and the factors are illustrated in Figure 5.12. The relationships were similar to previous models. The four models, shown in Table 5.20, indicated that the best fitting model was the one with only involvement as a predictor, $\chi^2(1) = 48.08$, p < .001. This model fit significantly better than the model with random effects removed, $\chi^2(2) = 695.76$, p < .001, *clm* model AIC = 7892.60, *clmm* model AIC = 7200.80.

Table 5.19

Polychoric Correlations Between Factors and Comprehensibility

Factor	Polychoric Correlation
Assuredness	.60 [.57, .64]
Involvement	.61 [.59, .64]
Positivity	.53 [.50, .56]

Figure 5.12
Relationships Between Factors and Comprehensibility

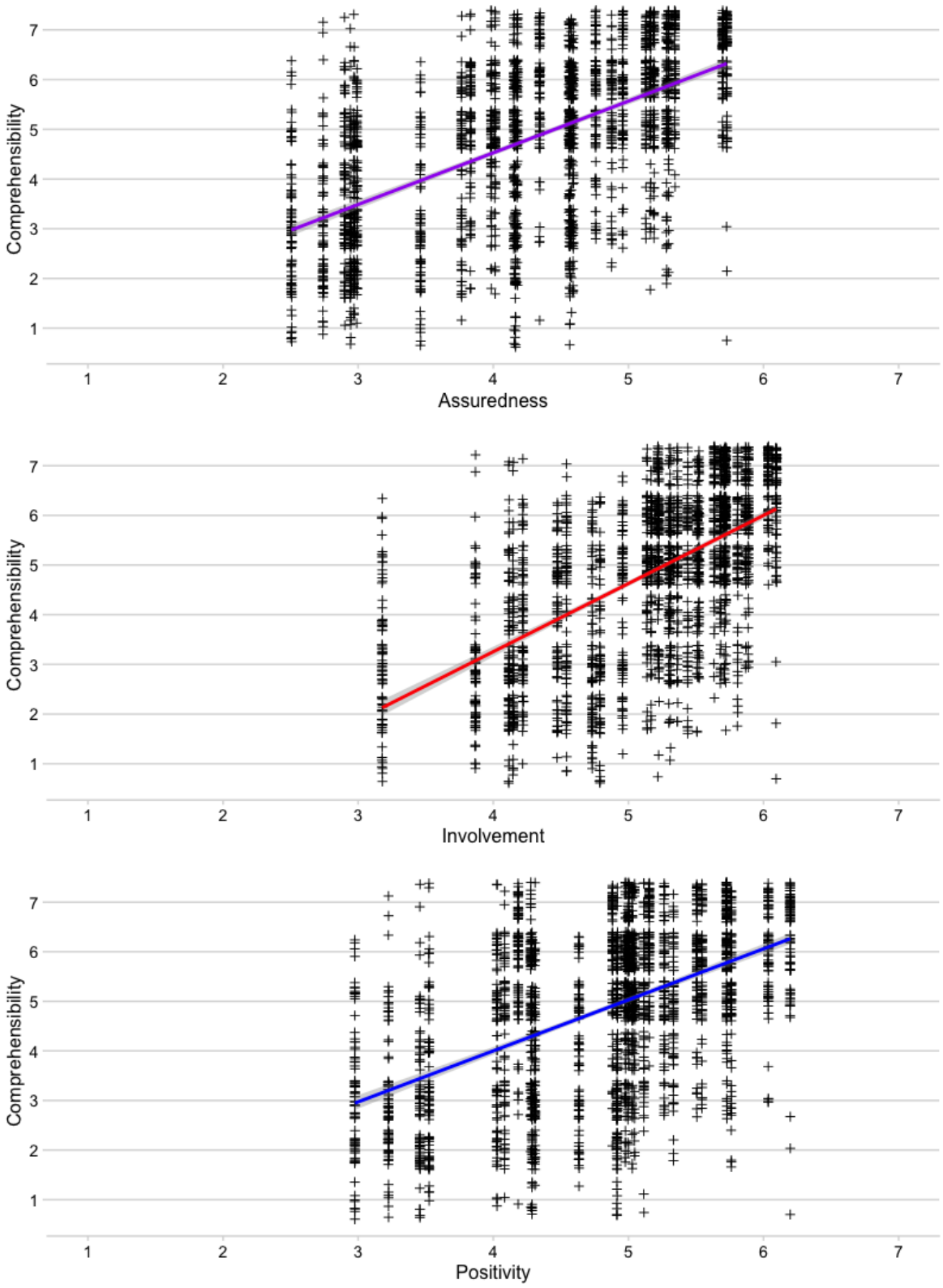


Table 5.20
Model Comparisons for Comprehensibility

Model	AIC	χ^2	*df*	*p*
Null Model	7246.90			
Model 1	7200.80	48.08	1	< .001
Model 2	7196.80	5.99	1	.014
Model 3	7198.60	0.22	1	.64

Note. Adjusted α = .0125.

The final model in Table 5.21 showed that involvement significantly predicted comprehensibility scores, β = 2.53, p < .001. The variance in this model was larger than in previous models, raters = 1.17, samples = 0.57. Involvement explained 25% of the variance in comprehensibility scores, Nagelkerke's Pseudo R^2 = .25.

Table 5.21
Final Comprehensibility Model

Coefficients	β	95% CI	*SE*	*z*	*p*	OR	95% CI
Involvement	2.53	[2.07, 2.99]	0.24	10.71	< .001	12.57	[7.90, 19.98]

Random effects		
Groups	Variance	SD
Raters	1.17	1.08
Samples	0.57	0.73

Summary

This chapter has considered two aspects of the dataset: structural integrity and relationships amongst rated variables. I first described how the dataset was cleaned to ensure the ratings were as reliable as possible prior to analysis. Of the 99 samples, I removed 16 due to low reliability, misfit, multivariate outliers, or a combination of these issues. I then checked the integrity of the dataset prior to analysis using descriptive statistics and Rasch measurement. The scales and samples functioned appropriately without misfit or erratic behavior. The raters, as expected, showed limited consistency and agreement, yet despite their lack of training, were able to assign scores that were consistent within reason.

I then considered interrelationships amongst the 14 rated variables of interest to RQ1. I found that all variables correlated, and some relationships were stronger. Four factors emerged from exploratory factor

analysis: Language, assuredness, involvement, and positivity. Using ordinal mixed-effects regression, these factor scores had different relationships with language proficiency outcomes. Assuredness alone was found to predict changes in fluency and grammar scores at a fairly high degree. Involvement, on the other hand, was the sole predictor for comprehensibility, which also showed a strong relationship. All three factor scores—assuredness, involvement, and positivity—predicted vocabulary, but only assuredness and involvement played a substantial role in predicting this outcome. Although rater effects such as the halo effect are undoubtedly part of the reason for these relationships, the results show that affect, in particular assuredness and involvement, may be tightly bound to language proficiency outcomes, while positive affect less strongly related.

These results describe relationships between variables observed by the raters in the study, which are subjective and prone to human error. In the next chapter, I turn to externally measured variables that are closely tied to nonverbal behavior. These measures were produced by a computer vision, machine learning algorithm, and will lend insight into whether omnibus indices of behavior also relate to proficiency outcomes.

CHAPTER 6: NONVERBAL BEHAVIOR AND LANGUAGE PROFICIENCY

While the rating data explored in Chapter 5 reveal interesting trends and patterns, one major limitation is that all variables were observed by the raters themselves, thus being prone to halo effects across categories. Likewise, understanding nonverbal behavior through affect is rather indirect, as affect may be derived from verbal information as well. In this chapter, I will explore the effects of external measures of nonverbal behavior derived from iMotions Affectiva, an automated pattern recognition algorithm that uses computer vision to classify the probability of positive or negative valence, engagement (a measure of expressiveness), and attention (a measure of eye gaze and head turn towards the camera). The research questions guiding this chapter are as follows:

RQ2.1: Do objectively measured indices of nonverbal behavior predict language proficiency scores?

RQ2.2: Do nonverbal behaviors moderate scores differentially depending on the proficiency levels of test takers?

In this chapter, I will first describe the data resulting from iMotions. I will present the distributions of each measure to illustrate differences and similarities amongst speech samples. I will also present graphical information to illustrate these trends. Following this, I will report inferential analyses of variables on the language proficiency measures.

Description and distribution of variables

Engagement

I first calculated mean scores and standard deviations for each of engagement, valence, and attention from the iMotions output files for the 30 samples, listed in Table 6.1. As can be seen in this table and Figure 6.1, mean engagement ranged from 5.15 (Sample 29; an overall low probability of exhibiting facial expressions) to 62.18 (Sample 23; a relatively high probability of exhibiting facial expressions). Figure 6.2 illustrates differences in engagement using an anonymized, cartoonized image of samples 23 and 29. Again, the actual video recordings of the test takers were used in the rating design, but for illustrative purposes only, I produced cartoonized images to protect the test takers' identities.

Table 6.1
iMotions Means and SDs by Sample

	Engagement		Valence		Attention	
Sample	*M*	*SD*	*M*	*SD*	*M*	*SD*
S01	15.50	24.08	-4.41	12.91	98.21	0.59
S02	11.09	20.54	-3.42	10.50	61.17	46.74
S03	7.21	17.19	-2.18	12.69	96.40	7.15
S04	50.50	39.55	33.38	44.54	97.93	2.59
S05	46.55	44.96	2.28	42.65	92.16	19.42
S06	11.60	24.60	-1.09	19.02	97.13	2.28
S07	22.44	31.42	-3.65	26.41	96.06	7.76
S08	42.16	41.61	-32.07	38.75	85.72	20.94
S09	11.45	18.09	-0.18	1.70	97.85	1.36
S10	19.71	28.87	-7.64	13.34	97.03	1.99
S11	7.95	16.00	-2.46	9.98	95.67	4.43
S12	19.24	28.09	2.21	16.26	97.10	1.24
S13	17.69	26.49	4.05	16.53	96.18	8.64
S14	56.84	37.36	40.22	39.26	96.01	6.76
S15	46.08	39.38	23.48	43.13	93.85	13.22
S16	10.15	19.14	-0.84	4.45	92.00	19.26
S17	19.77	25.30	-6.38	18.55	84.55	31.66
S18	25.16	37.91	13.88	33.55	93.60	9.68
S19	43.74	40.70	30.56	36.83	92.11	7.75
S20	16.61	28.21	0.24	29.32	60.24	45.85
S21	45.16	36.39	19.97	49.19	96.83	3.06
S22	23.71	33.39	10.16	28.72	97.15	1.46
S23	62.18	38.10	-1.33	11.46	96.86	3.10
S24	10.22	18.93	3.92	12.14	80.22	36.84
S25	39.31	40.20	12.31	32.16	80.33	22.40
S26	59.09	39.32	30.93	40.40	95.81	6.18
S27	19.13	31.44	6.71	27.73	74.82	38.17
S28	32.47	36.78	6.41	22.84	97.34	2.82
S29	5.15	11.78	-0.47	5.91	97.97	0.88
S30	30.28	30.91	0.19	17.45	93.78	20.17

Figure 6.1
Distribution of Engagement Means and SDs by Participant

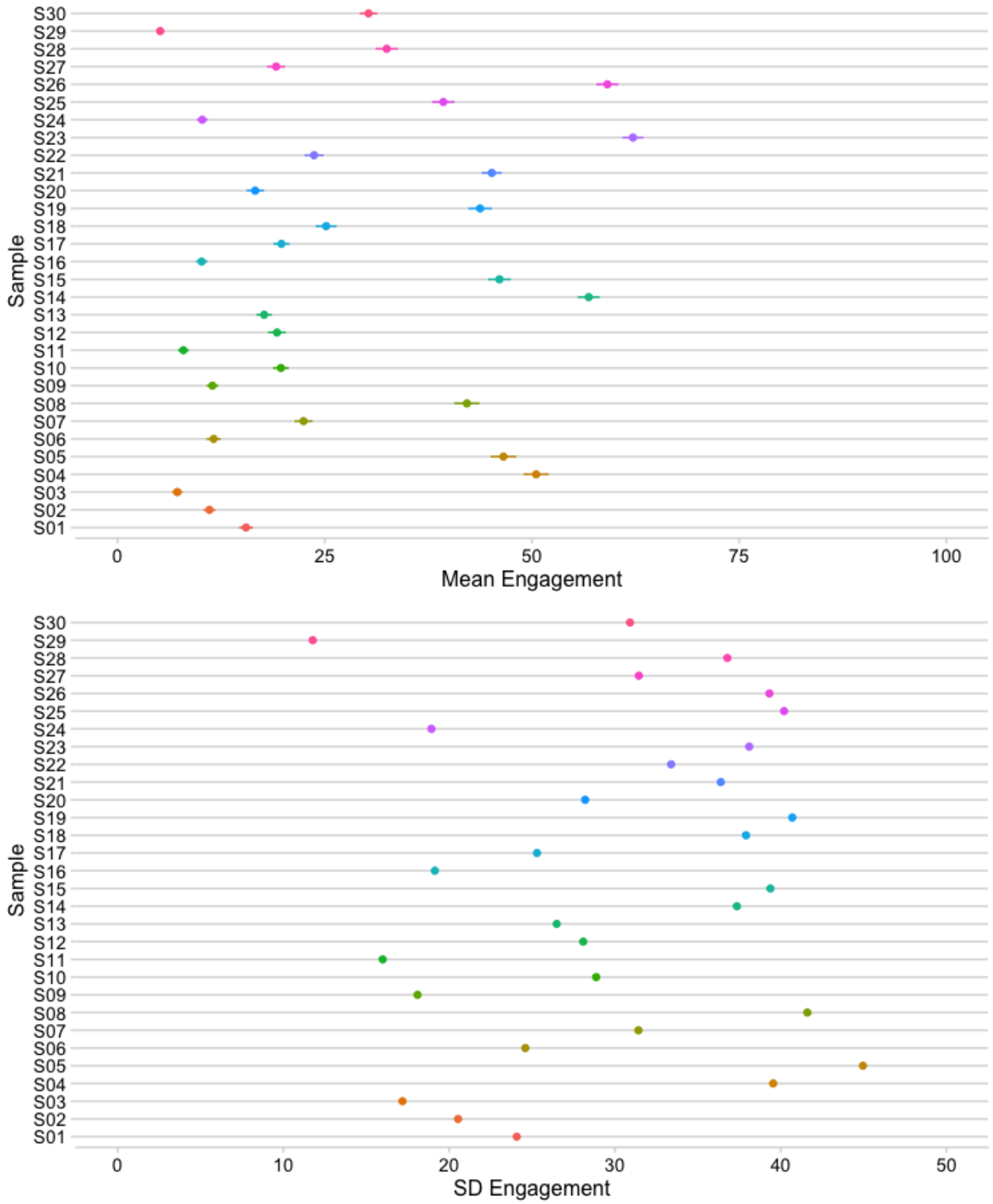
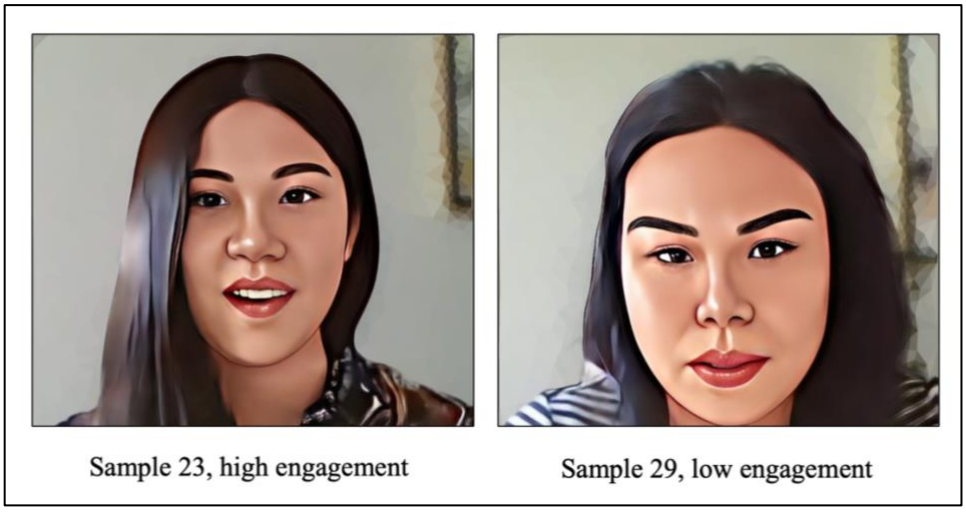


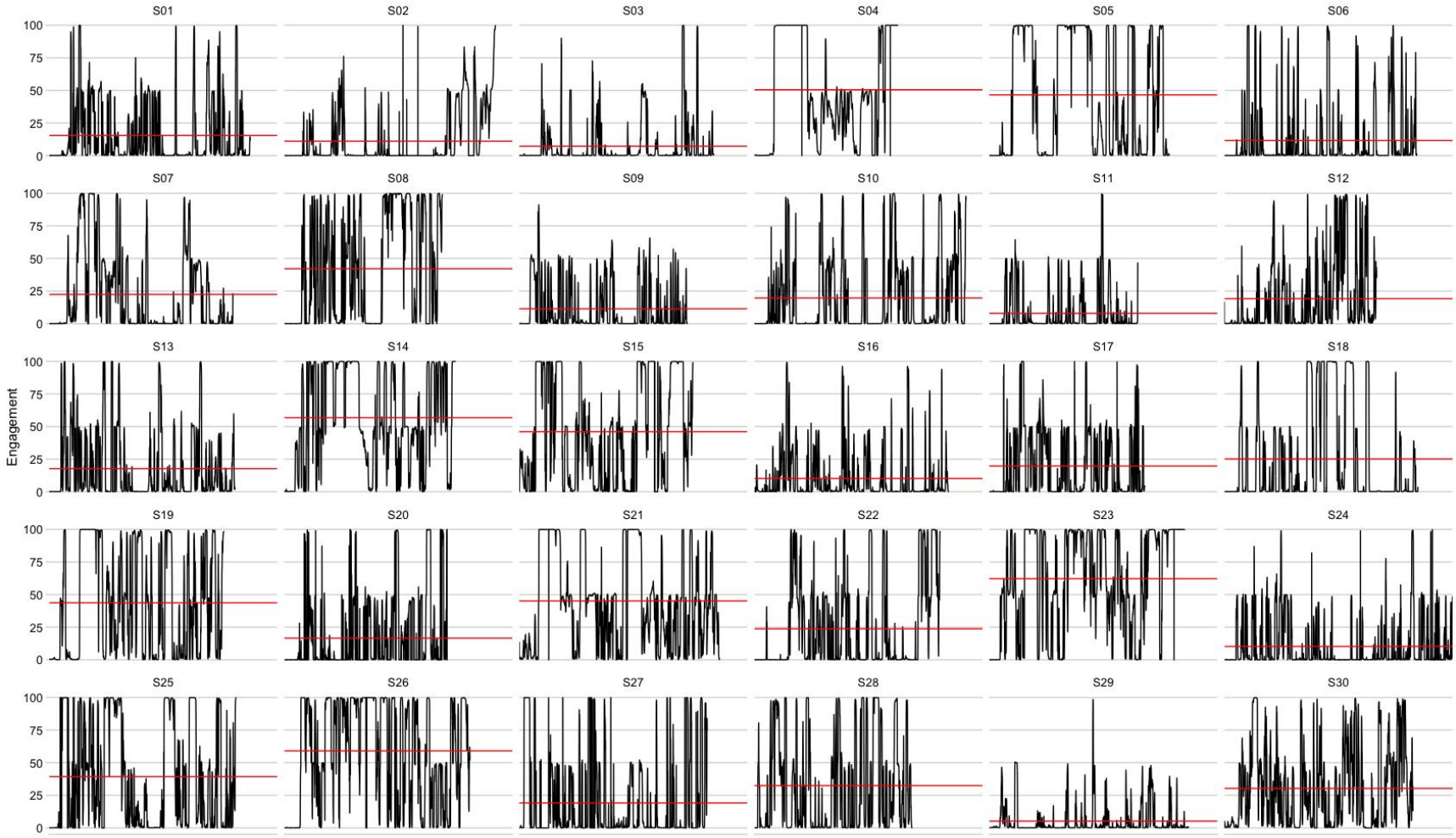
Figure 6.2
Illustration of High and Low Engagement



The standard deviations of engagement were quite large. This suggests that test takers in the speech samples varied substantially in their (probabilistic) appearance of facial expressions. Standard deviations ranged from a low of 11.78 (sample 29) to 44.96 (sample 5). Figure 6.3 visualizes these changes through a time series graph of the measurements of engagement during each speech sample. Here, the differences between samples 23 and 29 are even more striking. Sample 23 retained a high probability of facial expressiveness, especially during the middle of the test. However, despite a few peaks, the detection of facial expressiveness in Sample 29 was rather flat. The calculations of mean engagement and its standard deviation did not, however, correlate strongly with baseline proficiency level (.10 and .01 respectively).

Figure 6.3

Time Course Data for Engagement



Note. Red bar indicates mean engagement across entire sample

Valence

The mean scores and distribution of valence were markedly distinct from engagement. Mean scores, ranging from -100 to 100, did not generally deviate substantially from 0. These scores can be seen in Table 6.1 and visualized in Figure 6.4. The lowest mean valence score was -32.08 (Sample 8, *SD* = 38.75) while the highest valence score was 40.22 (Sample 14, *SD* = 39.26). Anonymized, cartoonized images of these two samples are presented in Figure 6.5. Standard deviations also varied widely, with Sample 21 exhibiting the most variance in valence (*M* = 19.97, *SD* = 49.19), and others, such as Sample 9, showing the least (*M* = -0.18, *SD* = 1.70). These trends are also visible in the time series graphs in Figure 6.6. These data suggest that most test takers did not exhibit strong evidence of positive or negative emotions throughout the test, which can be seen in examples such as Sample 9. “Flatline” time courses suggest a somewhat unchanging emotional appearance.

Figure 6.4
Distribution of Valence Means and SDs by Participant

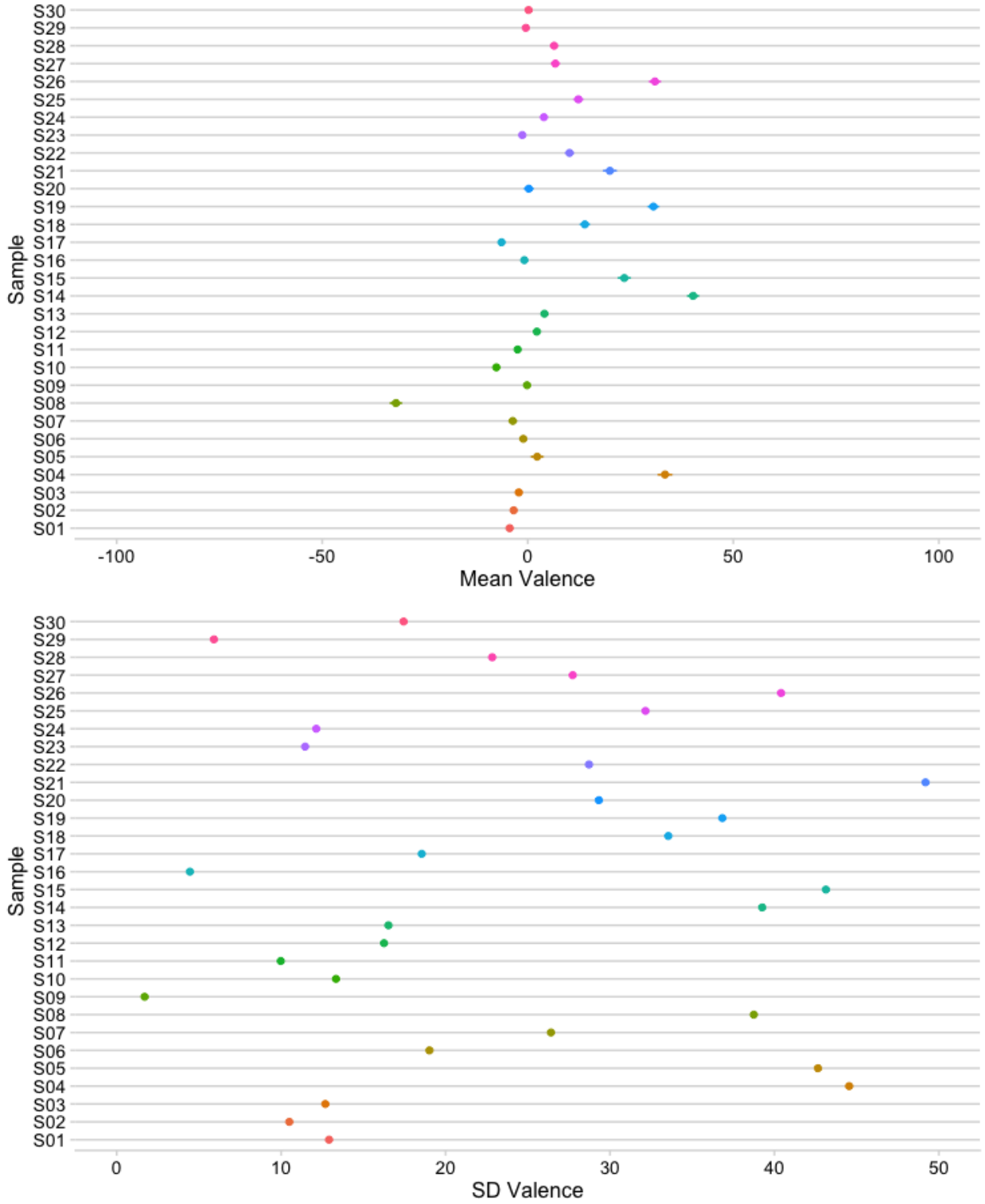
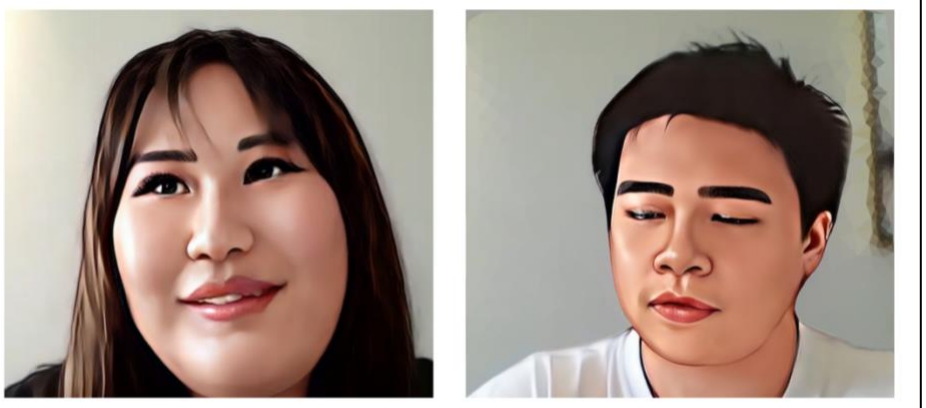


Figure 6.5
Illustration of High and Low Valence

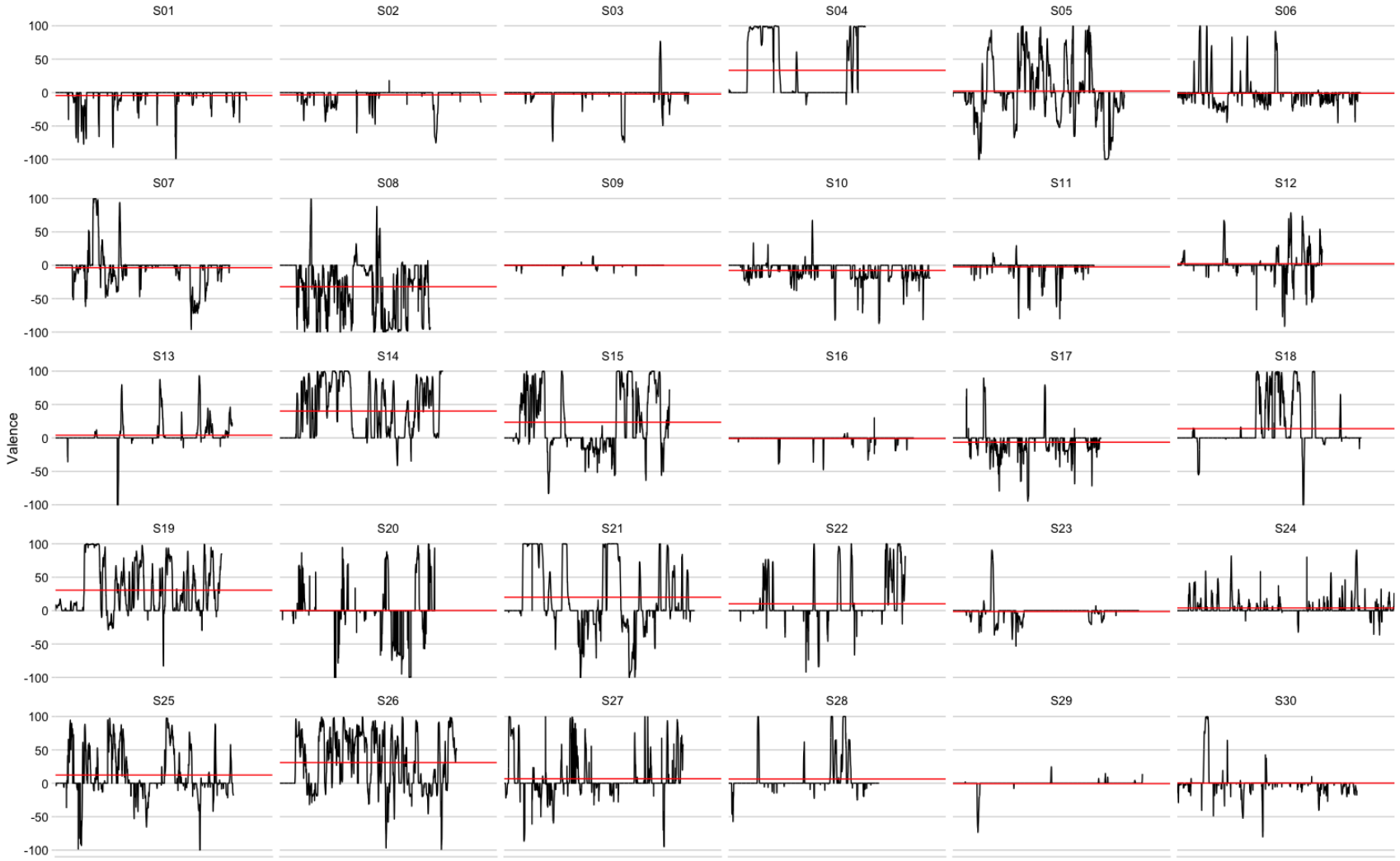


Sample 14, high valence

Sample 8, low valence

Figure 6.6

Time Course Data for Valence



Note. Red bar indicates mean valence across entire sample

Attention

Attention scores also varied in their distribution from engagement and valence scores. As seen in Table 6.1 and visualized in Figure 6.7, attention tended to be high across the samples, and variance tended to be low with the exception of a small number of samples. The sample with the highest mean attention was Sample 1, with a mean probability of 98.21 (*SD* = 0.59). This individual moved his head very little, especially side to side movements. This sample also spent much of the time looking towards the camera. When gaze was broken, his head stayed fixed towards the camera. The sample with the lowest mean attention was Sample 20, with a mean probability of 60.24 (*SD* = 45.85). This test taker frequently turned her head to one side or the other, and also frequently broke gaze with the interlocutor. Figure 6.9 shows how attention changed throughout the speech samples. Samples 20 and 24, for example, appeared to shift their attention frequently, while Samples 1 and 29 tended to hold their attention more focused.

Figure 6.7

Distribution of Attention Means and SDs by Participant

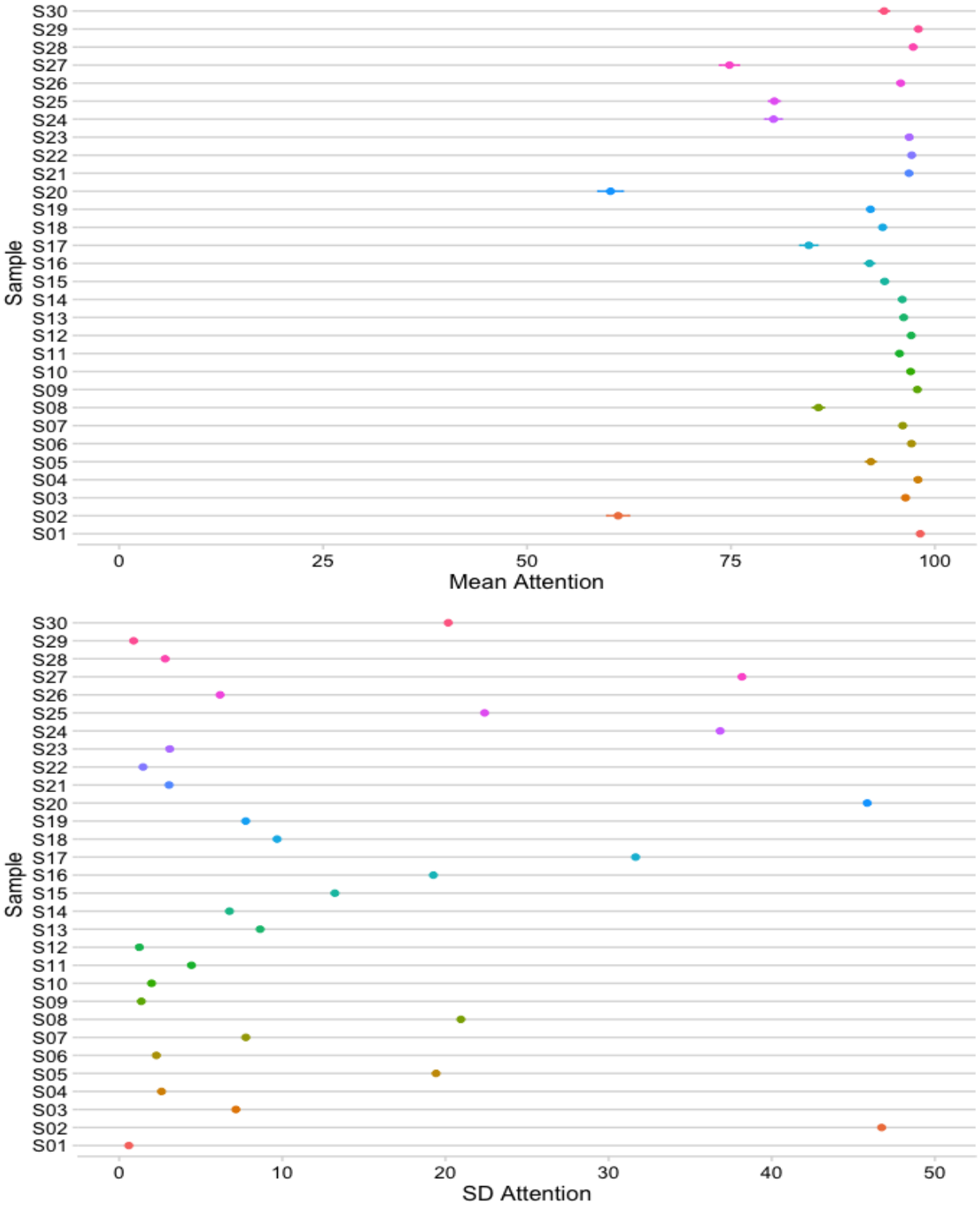
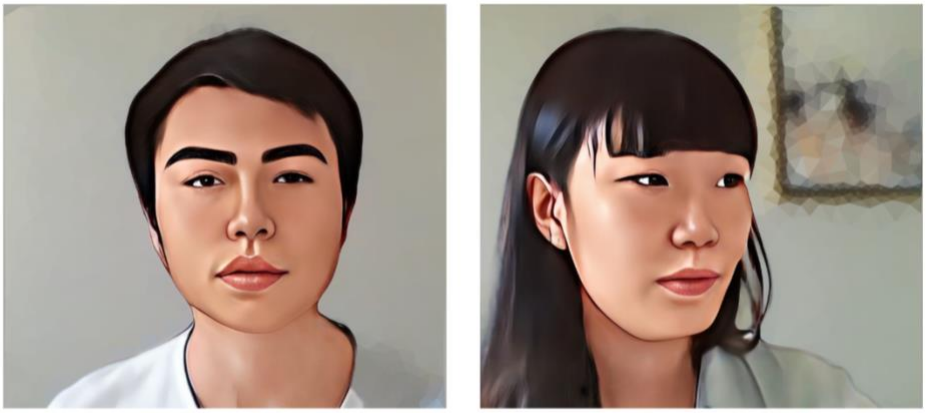


Figure 6.8
Illustration of High and Low Attention

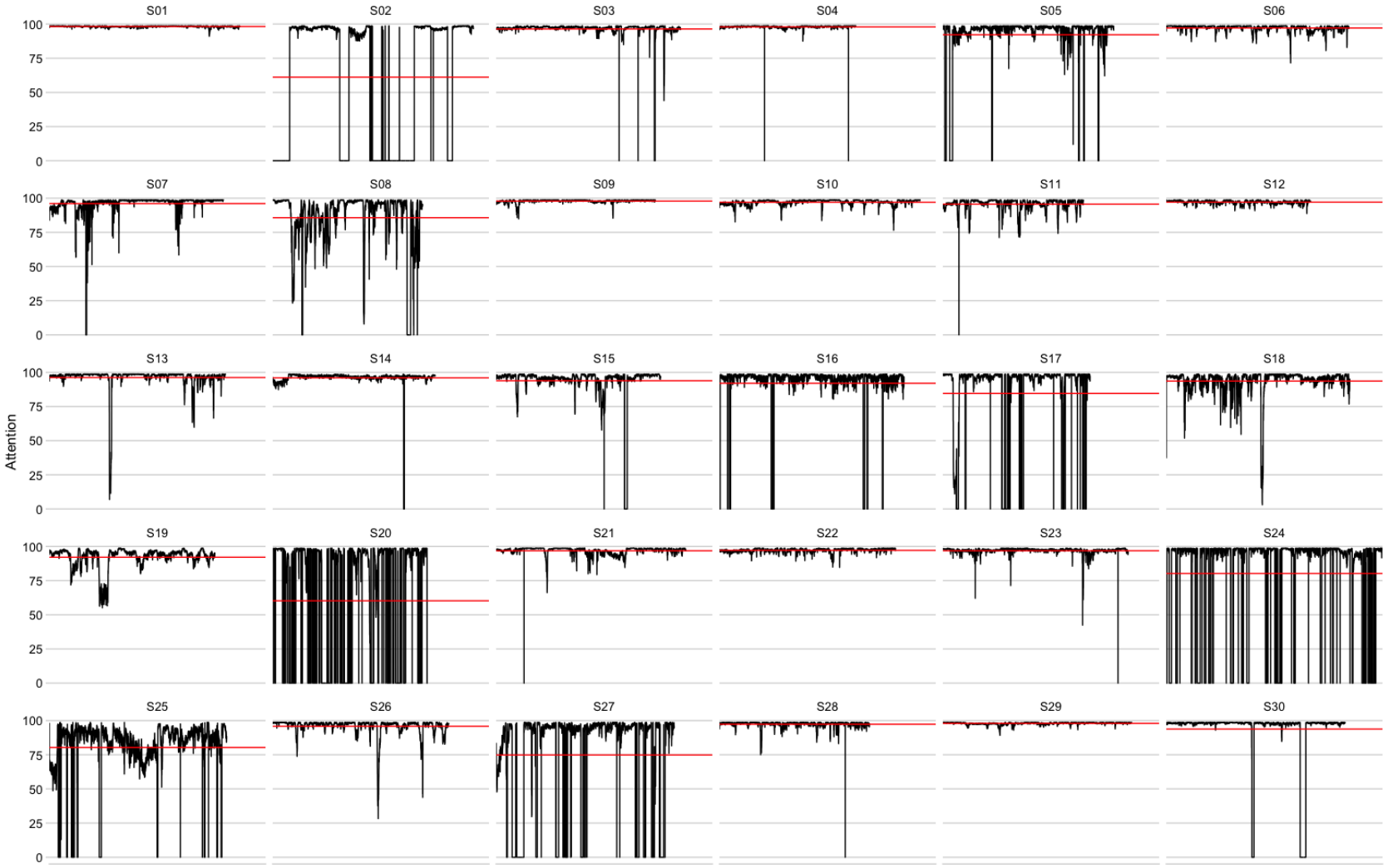


Sample 1, high attention

Sample 20, low attention

Figure 6.9

Time Course Data for Attention



Note. Red bar indicates mean attention across entire sample.

Correlations with affect scores

Finally, I calculated Pearson correlations between the three iMotions sets of variables and the factor scores from Chapter 5, shown in Table 6.2. These correlations showed that the iMotions variables were somewhat interrelated. Engagement and valence correlated at .54, which is a medium correlation. This is logical, as showing positive valence is dependent on showing a certain amount of expressiveness. Engagement also correlated with attention at .20, and valence and attention correlated at .19. While related, these variables did appear to be measuring different aspects of performance. The means and standard deviations also correlated quite strongly. Mean engagement correlated with its standard deviation at .87, which indicates that more overall expressive individuals were more likely to vary the intensity of their expressiveness throughout the sample, which is logical. Mean valence correlated with its standard deviation at .56, which may be interpreted similarly. Mean attention correlated with its standard deviation at -.94, which indicates that individuals with lower mean attention had more variance in how they established and broke attention with the interlocutor, while individuals with higher mean attention were less likely to break their attention frequently.

Table 6.2
Pearson Correlations Amongst iMotions Variables and Affective Factors

	1	2	3	4	5	6	7	8
1. Engagement (*M*)								
2. Valence (*M*)	.54							
3. Attention (*M*)	.20	.19						
4. Engagement (*SD*)	.87	.42	.08					
5. Valence (*SD*)	.74	.56	(-.001)	.85				
6. Attention (*SD*)	-.19	-.22	-.94	-.08	(-.01)			
7. Assuredness	.05	.11	-.02	-.02	-.13	.12		
8. Involvement	.21	.24	.09	.18	.09	(.01)	.92	
9. Positivity	.48	.39	.06	.43	.41	(.01)	.79	.91

Note. All correlations are significant except those in (parentheses).

Regarding the correlations between iMotions variables and the factor scores, there was some indication that mean engagement and positivity were measuring similar attributes, as they correlated at .48. Note, however, that expressiveness was one of the variables that factored into positivity. Mean valence, which would logically have correlated with positivity more, correlated somewhat less at .39. Involvement

had weak correlations with engagement (.21) and valence (.24), while the remaining correlations were quite negligible. The correlations with the standard deviations followed similar patterns.

Nonverbal behavior and facets of proficiency

Fluency

Correlations. Table 6.3 shows the polychoric correlations between each iMotions variable and fluency. The only significant correlation between the mean behavior scores and fluency was valence, which correlated at .12. This indicates that there was a small positive relationship between positive valence and fluency level. The standard deviations of valence and attention also correlated with fluency at -.11 and .11, respectively. These correlations indicate a negative relationship between variance in valence (e.g., alternating between positive and negative valence) and fluency, and a positive relationship between higher attentional variance and fluency. These correlations are illustrated in

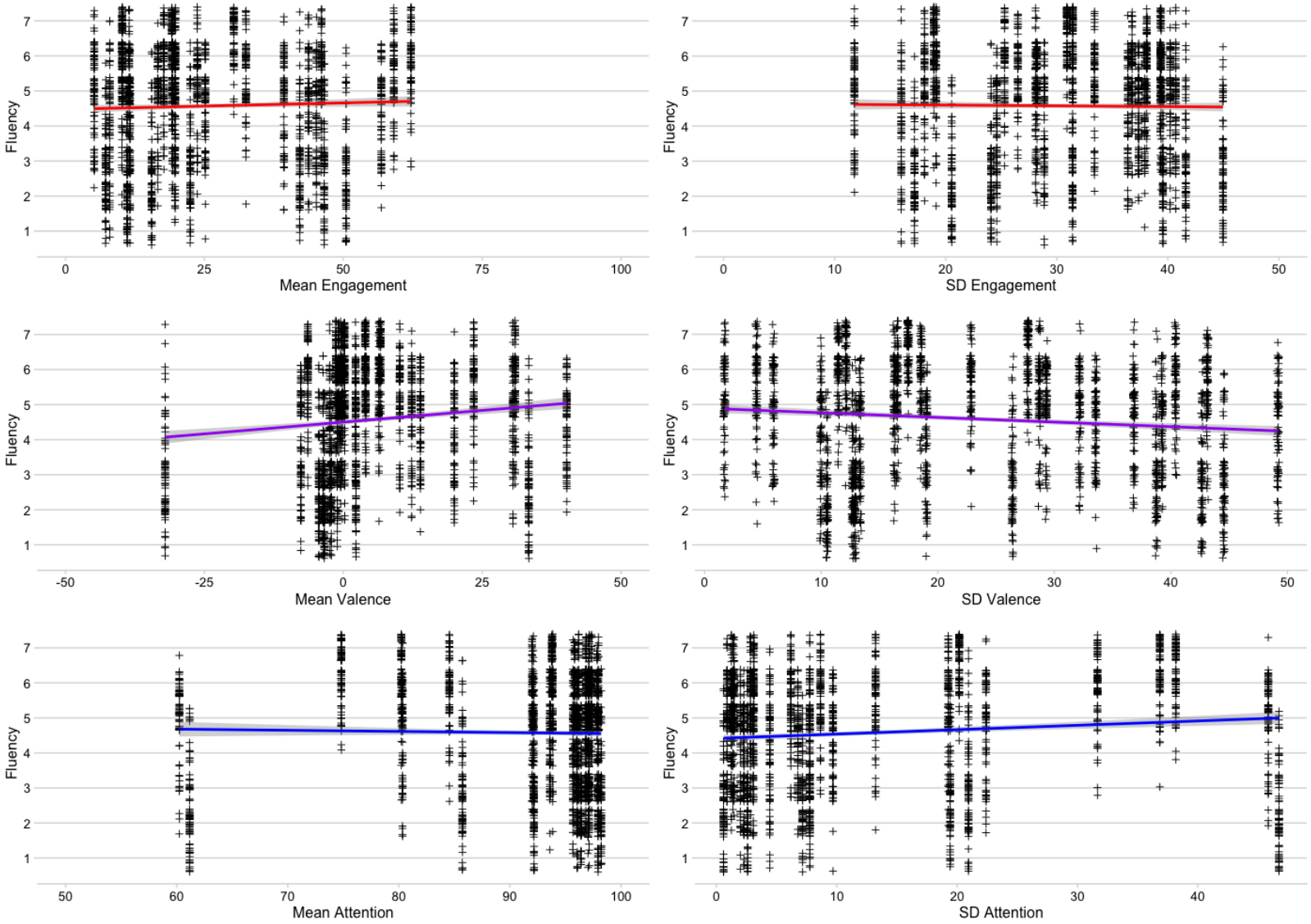
Figure 6.10 as regression lines. Similar to Chapter 5, in these figures, vertical lines represent the distribution of scores for each test taker.

Table 6.3
Polychoric Correlations with Fluency

Variable	Mean	*SD*
Engagement	.04 [0, .08]	-.01 [-.06, .03]
Valence	.12 [.08, .16]	-.11 [-.15, -.07]
Attention	-.02 [-.07, .03]	.11 [.07, .15]

Figure 6.10

Relationships Between Nonverbal Measures and Fluency



Regression of mean predictors. Factors were entered in the model based on the correlations with fluency listed in Table 6.3. The five models, shown in Table 6.4, indicated that the best fitting model was the interaction model. The interaction model, presented in **Error! Reference source not found.**, fit significantly better than the other models, $\chi^2(4) = 50.05$, $p < .001$. This model also fit significantly better than the model with random effects removed, $\chi^2(2) = 712$, $p < .001$. The only significant predictor in this model was base proficiency (the scaled IELTS scores), which had a sizeable effect on fluency ($\beta = 2.27$, odds ratio = 9.66). This model, however, only explained minimal variance in the model, Nagelkerke's Pseudo $R^2 = .02$.

Table 6.4
Model Comparisons for Fluency (Means)

Model	AIC	χ^2	*df*	*p*
Null Model	7016.70			
Model 1	7018.10	0.56	1	.45
Model 2	7020.10	0.05	1	.83
Model 3	7021.90	0.12	1	.72
Interaction model	6980.00	50.05	4	< .001

Table 6.5
Interactions Between Base Proficiency and Mean Behavioral Indices on Fluency

Coefficients	β	95% CI	*SE*	*z*	*p*	OR	95% CI
Base proficiency	2.27	[0.76, 3.78]	0.77	2.95	.003	9.66	[2.14, 43.60]
Valence	0.04	[-0.03, 0.11]	0.04	1.05	.30	1.04	[0.97, 1.10]
Engagement	-0.04	[-0.10, 0.01]	0.03	-1.45	.15	0.96	[0.91, 1.00]
Attention	-0.07	[-0.004, 0.14]	0.04	1.85	.06	1.07	[0.996, 1.20]
Val:Prof	-0.01	[-0.03, 0.007]	-0.01	-1.25	.21	0.99	[0.97, 1.00]
Eng:Prof	0.01	[-0.003, 0.02]	0.01	1.53	.12	1.01	[0.997, 1.00]
Att:Prof	-0.02	[-0.03, 0.001]	0.02	-1.83	.07	0.98	[0.97, 1.00]

Random effects		
Groups	Variance	SD
Raters	0.97	0.99
Samples	0.77	0.88

Note. Adjusted $\alpha = .0125$.

Regression of predictor standard deviations. Similarly, predictor standard deviations were entered in this secondary model based on the absolute value of correlations with fluency listed in Table 6.3. Because the absolute values of valence and attention were equivalent, I entered valence in the first model

for comparability with the models of mean indices. The five models, including the interaction model shown in Table 6.6, indicated that the best fitting model was the interaction model, $\chi^2(4) = 53.48$, $p < .001$. This model also fit significantly better than the model with random effects removed, $\chi^2(2) = 647$, $p < .001$. This model, presented in Table 6.7, contrasted with the mean model in that one interaction term, attention with base proficiency, was significant $\beta = 0.02$, $p = .003$, with a very small effect size (odds ratio = 1.02). The main effect of attention was also significant, $\beta = -0.06$, $p = .01$, but I will not interpret the main effect given the significant interaction term. This model explained minimal variance in the outcome, Nagelkerke's Pseudo $R^2 = .03$.

Table 6.6
Model Comparisons for Fluency (SDs)

Model	AIC	χ^2	*df*	*p*
Null Model	7017			
Model 1	7018	0.79	1	.37
Model 2	7018	1.40	1	.24
Model 3	7019	1.38	1	.24
Interaction model	6974	53.48	4	< .001

Table 6.7
Interactions Between Base Proficiency and Behavioral Index SDs on Fluency

Predictors	β	95% CI	*SE*	*z*	*p*	OR	95% CI
Base proficiency	0.76	[0.07, 1.46]	0.35	2.15	.03	2.15	[1.07, 4.30]
Valence	-0.08	[-0.25, 0.08]	0.09	-0.95	.34	0.92	[0.78, 1.09]
Attention	-0.06	[-0.11, -0.01]	0.26	-2.45	.01	0.94	[0.89, 0.99]
Engagement	0.09	[-0.15, 0.33]	0.12	0.73	.46	1.09	[0.86, 1.38]
Val:Prof	0.01	[-0.03, 0.04]	0.01	0.35	.73	1.01	[0.97, 1.04
Att:Prof	0.02	[0.01, 0.03]	0.01	3.04	.002	1.02	[1.01, 1.03]
Eng:Prof	-0.005	[-0.05, 0.04]	-0.001	-0.20	.84	1.00	[0.95, 1.04]

Random effects

Groups	Variance	SD
Raters	0.97	0.99
Samples	0.61	0.78

Note. Adjusted $\alpha = .0125$.

To explore the interactions, I dichotomized proficiency and reran the simple ordinal model only using the interaction terms *fluency~sd_Attention*proficiency*, as detailed in the methods section. Significance tests of the differences in the coefficients between the two proficiency groups for all 7 levels

of fluency score are presented in Table 6.8. The comparisons use the high proficiency group as the comparison group. They show that the impact of attention on the proficiency groups was not even for each score level. For example, at a score level of 3, the standard deviation of attention had less of an positive effect on the higher group than the lower group ($\beta = -0.17$). The direction of impact switched between scores of 4 and 5. At a score of 6, for example, the high group was more positively affected by the variance in attention than the lower group ($\beta = 0.25$). Figure 6.11 shows the probabilities of a particular score assignment according to the variance in attention. At lower levels of proficiency, greater attention corresponded with lower probabilities of reaching a score of 6 to 7. However, at higher levels of proficiency, the effect was the opposite, with greater variance in attention leading greater probabilities of scoring in higher brackets.

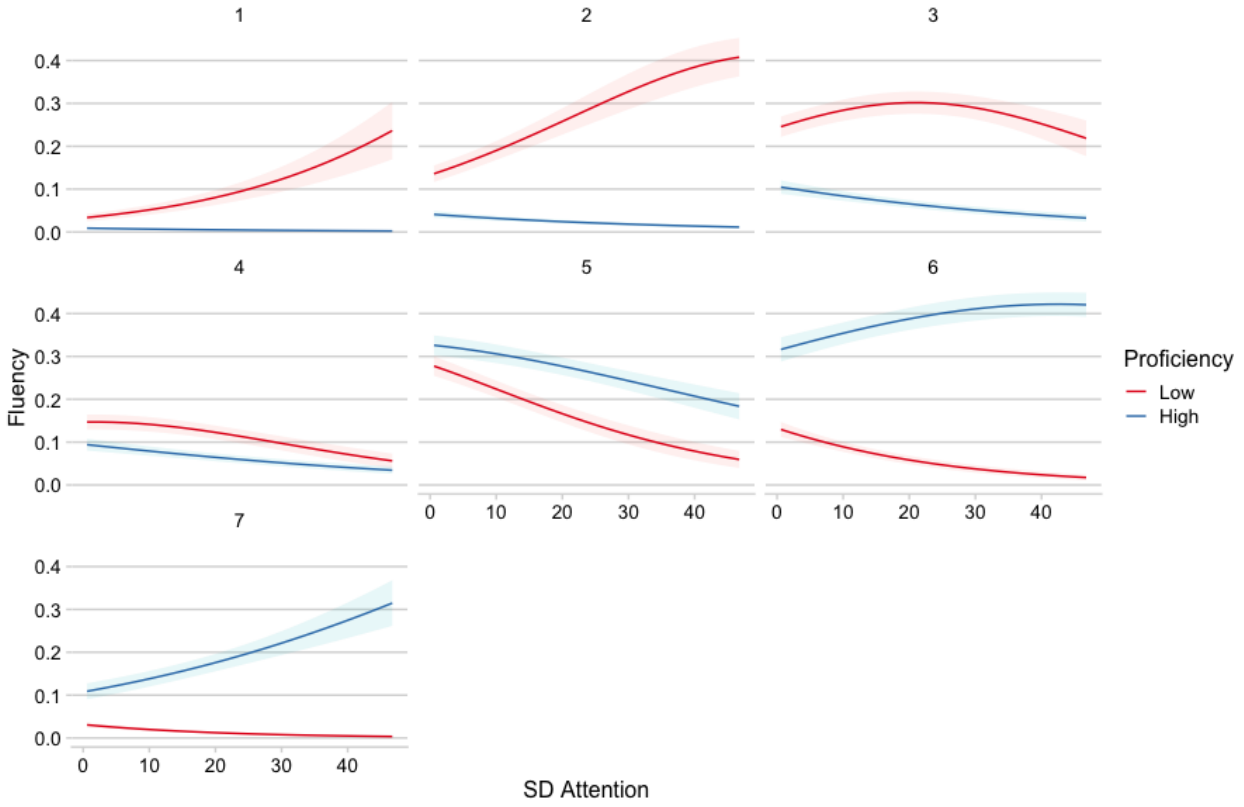
Table 6.8

Post hoc Comparisons for Fluency with SD Attention

Score level	β	95% CI	*SE*	*z*	*p*	OR	95% CI
1	-0.08	[-0.06, -0.09]	0.01	-9.33	< .001	0.93	[0.91, 0.94]
2	-0.18	[-0.16, -0.21]	0.01	-17.87	< .001	0.83	[0.81, 0.85]
3	-0.17	[-0.15, -0.19]	0.01	-16.92	< .001	0.84	[0.83, 0.86]
4	-0.04	[-0.03, -0.05]	0.005	-8.60	< .001	0.96	[0.95, 0.97]
5	0.08	[0.10, 0.06]	0.01	9.57	< .001	1.08	[1.07, 1.10]
6	0.25	[0.27, 0.23]	0.01	22.33	< .001	1.29	[1.26, 1.31]
7	0.14	[0.15, 0.12]	0.01	15.82	< .001	1.15	[1.13, 1.17]

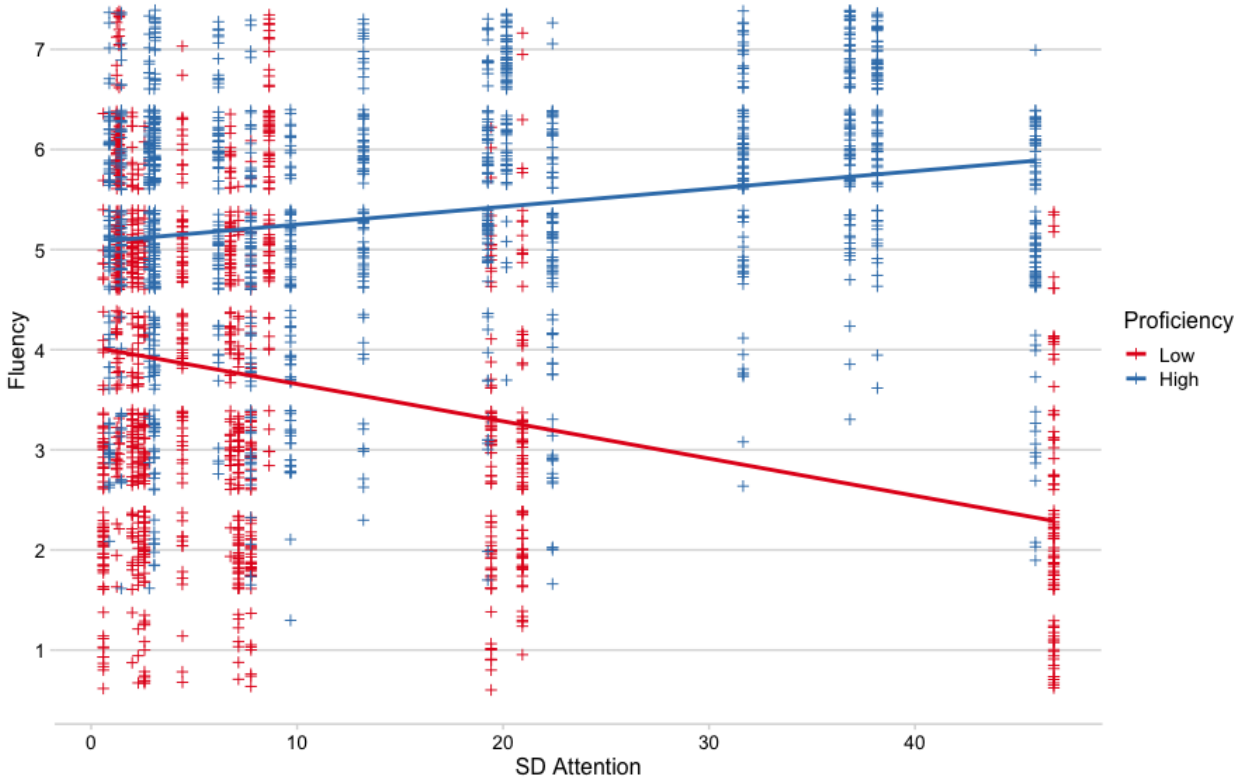
Note. Adjusted α = .00714 for 7 comparisons

Figure 6.11
Probability of Fluency Score Given SD Attention



Finally, I illustrate the effects of this interaction on the dataset as a whole, pictured in Figure 6.12. Although the effect size was in reality quite small, it can be seen that more varied attention related to differential outcomes for the two proficiency groups.

Figure 6.12
Visualization of Impact of SD Attention on Fluency



Vocabulary

Correlations. Table 6.9 displays the polychoric correlations between each behavioral index and vocabulary, while Figure 6.13 visualizes these relationships. These relationships were identical in direction to fluency, but slightly different in size. The relationship between vocabulary and mean valence was slightly weaker at .08 (rather than .12), as well as between vocabulary and the standard deviation of valence (-.15). Correlations with the standard deviation of attention were the same as with fluency (.11).

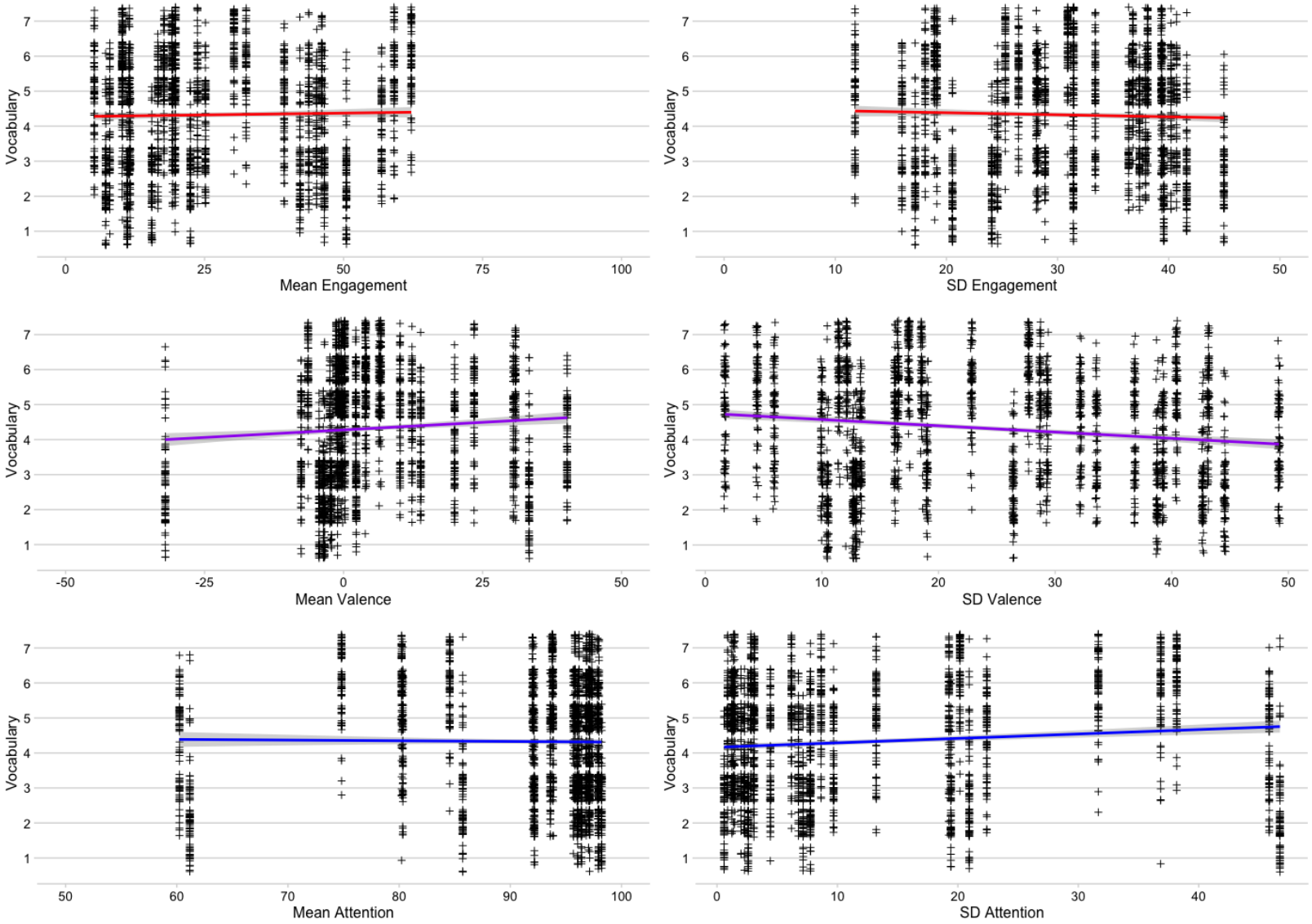
Table 6.9

Polychoric Correlations With Vocabulary

Variable	Mean	*SD*
Engagement	.02 [-.01, .06]	-.03 [-.07, .01]
Valence	.08 [.03, .12]	-.15 [-.19, -.11]
Attention	-.01 [-.06, .03]	.11 [.06, .15]

Figure 6.13

Relationships Between Nonverbal Measures and Vocabulary



Regressions of main predictors. Factors were entered in the model based on the correlations with vocabulary listed in Table 6.9. The results were nearly identical for fluency. The five-model comparison, shown in Table 6.10, indicated that the best fitting model was the interaction model. The interaction model, presented in Table 6.11, fit significantly better than the other models, $\chi^2(4) = 52.44$, $p < .001$. This model also fit significantly better than the model with random effects removed, $\chi^2\ (2) = 539.09$, $p < .001$. As with fluency, however, the only significant predictor in this model was base proficiency, which had a sizeable effect on vocabulary ($\beta = 1.90$, odds ratio = 6.70). This model only explained minimal variance in the model, Nagelkerke's Pseudo $R^2 = .02$.

Table 6.10
Model Comparisons for Vocabulary (Means)

Model	AIC	χ^2	*df*	*p*
Null Model	7325.30			
Model 1	7327.00	0.33	1	.56
Model 2	7328.90	0.03	1	.86
Model 3	7330.90	0.02	1	.90
Interaction model	7286.60	52.44	4	< .001

Table 6.11
Interactions Between Base Proficiency and Mean Behavioral Indices on Vocabulary

Coefficients	β	95% CI	*SE*	*z*	*p*	OR	95% CI
Base proficiency	1.90	[0.56, 3.24]	0.68	2.78	.005	6.70	[1.75, 25.47]
Valence	0.04	[-0.03, 0.10]	0.03	1.08	.28	1.04	[0.97, 1.11]
Engagement	-0.05	[-0.10, 0.01]	0.03	-1.72	.09	0.96	[0.91, 1.01]
Attention	0.06	[-0.001, 0.13]	0.03	1.93	.05	1.07	[1.00, 1.14]
Val:Prof	-0.01	[-0.03, 0.005]	0.01	-1.45	.15	0.99	[0.97, 1.00]
Eng:Prof	0.01	[-0.001, 0.02]	0.01	1.83	.07	1.01	[1.00, 1.02]
Att:Prof	-0.01	[-0.03, 0.003]	0.01	-1.62	.11	0.99	[0.97, 1.00]

Random effects

Groups	Variance	SD
Raters	0.65	0.81
Samples	0.60	0.77

Note. Adjusted $\alpha = .0125$.

Regression of predictor standard deviations. Similarly, predictor standard deviations were entered in this secondary model based on the absolute value of correlations with vocabulary listed in Table 6.9. The five models, shown in Table 6.12, indicated that the best fitting model was the interaction model,

$\chi^2(4) = 54.57$, $p < .001$. This model, shown in Table 6.13, also fit significantly better than the model with random effects removed, $\chi^2(2) = 647$, $p < .001$. This model, similar to that of fluency, had one interaction term, attention with base proficiency, which was significant, $\beta = 0.02$, $p = .003$, with a very small effect size (odds ratio = 1.02). This model explained minimal variance in the outcome, Nagelkerke's Pseudo $R^2 = .02$.

Table 6.12
Model Comparisons for Vocabulary (SDs)

Model	AIC	χ^2	*df*	*p*
Null Model	7325.30			
Model 1	7326.10	1.22	1	.27
Model 2	7327.40	0.65	1	.41
Model 3	7327.10	2.30	1	.13
Interaction model	7280.60	54.57	4	<.001

Table 6.13
Interactions Between Base Proficiency and Behavioral Index SDs on Vocabulary

Predictors	β	95% CI	*SE*	*z*	*p*	OR	95% CI
Base proficiency	0.69	[0.06, 1.31]	0.32	2.16	.03	1.99	[1.07, 3.70]
Valence	-0.08	[-0.23, 008]	0.08	-0.97	.33	0.93	[0.80, 1.08]
Attention	-0.06	[-0.10, -0.01]	0.02	-2.37	.02	0.95	[0.90, 0.99]
Engagement	0.07	[-0.18, 0.29]	0.11	0.69	.49	1.08	[0.87, 1.33]
Val:Prof	0.003	[-0.03, 0.03]	0.02	0.21	.83	1.00	[0.97, 1.03]
Att:Prof	0.01	[0.004, 0.02]	0.01	2.79	.005	1.01	[1.00, 1.02]
Eng:Prof	-0.001	[-0.04, 0.04]	0.02	-0.05	.96	1.00	[0.96, 1.04]

Random effects

Groups	Variance	SD
Raters	0.68	0.81
Samples	0.49	0.70

Note. Adjusted $\alpha = .0125$.

As for fluency, I calculated the differences in the coefficients and odds ratios between the proficiency groups by each vocabulary score, presented in Table 6.14. The comparisons show nearly identical trends with the fluency model, except that a score level of 4 showed no differences in the proficiency groups in how the standard deviation of attention impacted fluency. Figure 6.14 shows the probabilities of a particular score assignment according to the variance in attention, with analogous trends with fluency. This indicated that higher variance in attention, such as shifting gaze frequently, increased a

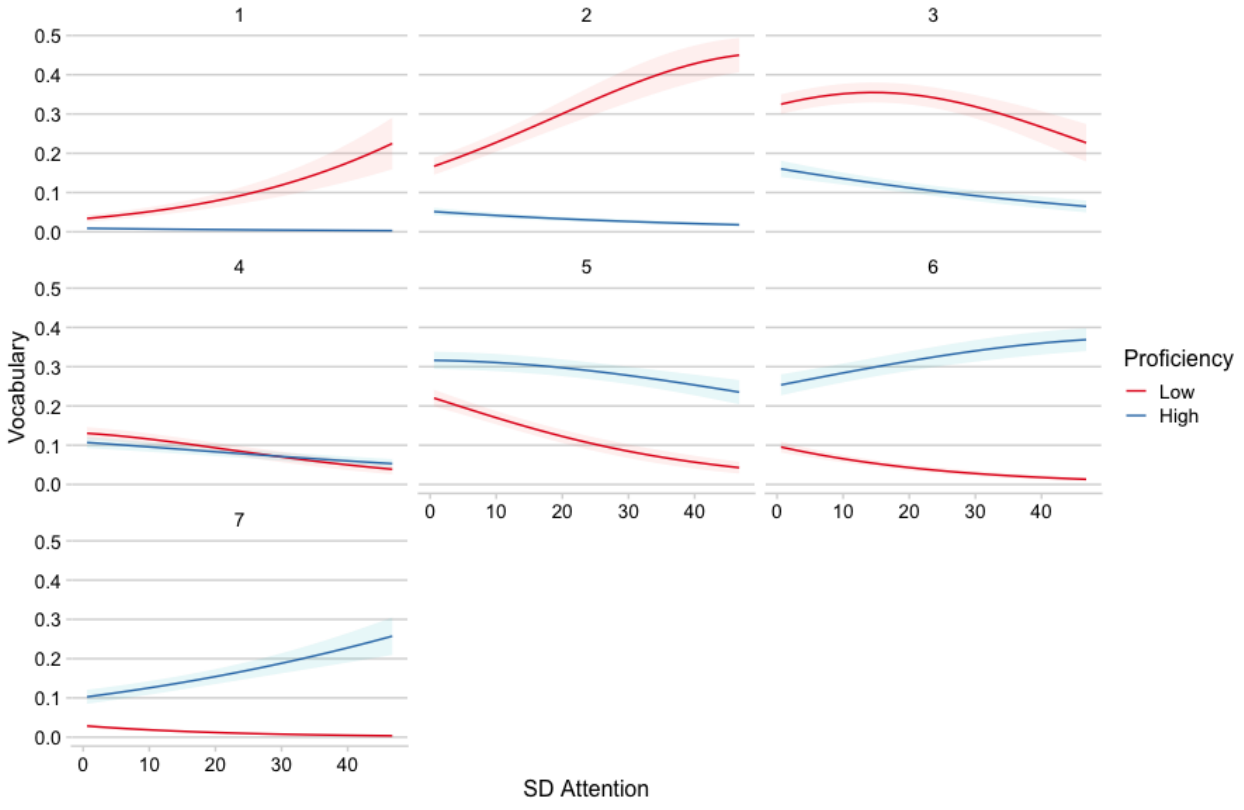
less proficient individual's chance of receiving a 1 or 2 and lowered their chances of being awarded higher scores.

Table 6.14
Post hoc Comparisons for Vocabulary with SD Attention

Score level	β	95% CI	*SE*	*z*	*p*	OR	95% CI
1	-0.07	[-0.06, -0.09]	0.01	-9.24	<.001	0.93	[0.91, 0.94]
2	-0.21	[-0.19, -0.23]	0.01	-19.48	<.001	0.81	[0.79, 0.83]
3	-0.17	[-0.15, -0.19]	0.01	-15.96	<.001	0.85	[0.83, 0.86]
4	-0.01	[-0.001, -0.02]	0.004	-2.15	.03	0.99	[0.98, 1.00]
5	0.13	[0.14, 0.11]	0.01	15.32	<.001	1.13	[1.12, 1.15]
6	0.21	[0.23, 0.19]	0,01	20.32	<.001	1.23	[1.21, 1.26]
7	0.12	[0.14, 0.11]	0.01	14.85	<.001	1.13	[1.11, 1.15]

Note. Adjusted α = .00714 for 7 comparisons

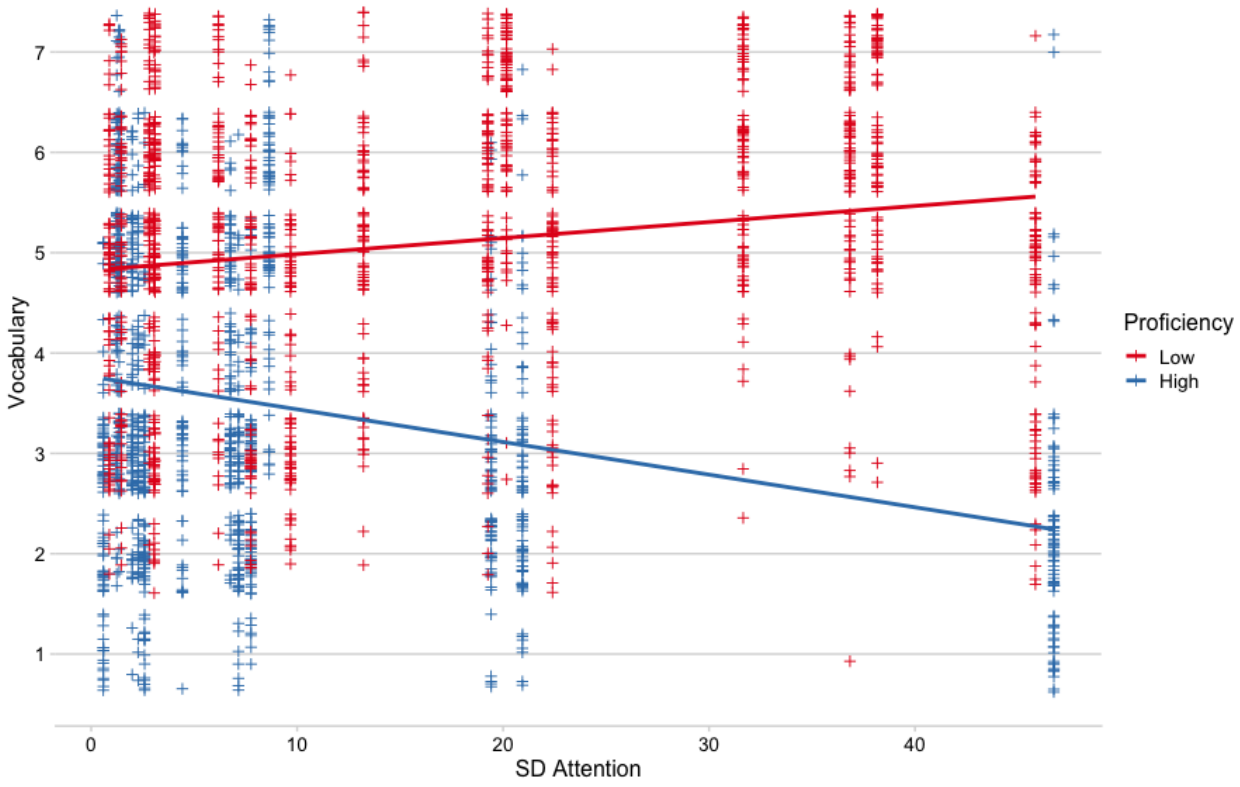
Figure 6.14
Probability of Vocabulary Score Given SD Attention



I illustrate the effects of this interaction on the dataset as a whole, pictured in Figure 6.15. As with fluency, it can be seen that more varied attention related to differential outcomes for the two proficiency

groups.

Figure 6.15
Visualization of Impact of SD Attention on Vocabulary



Grammar

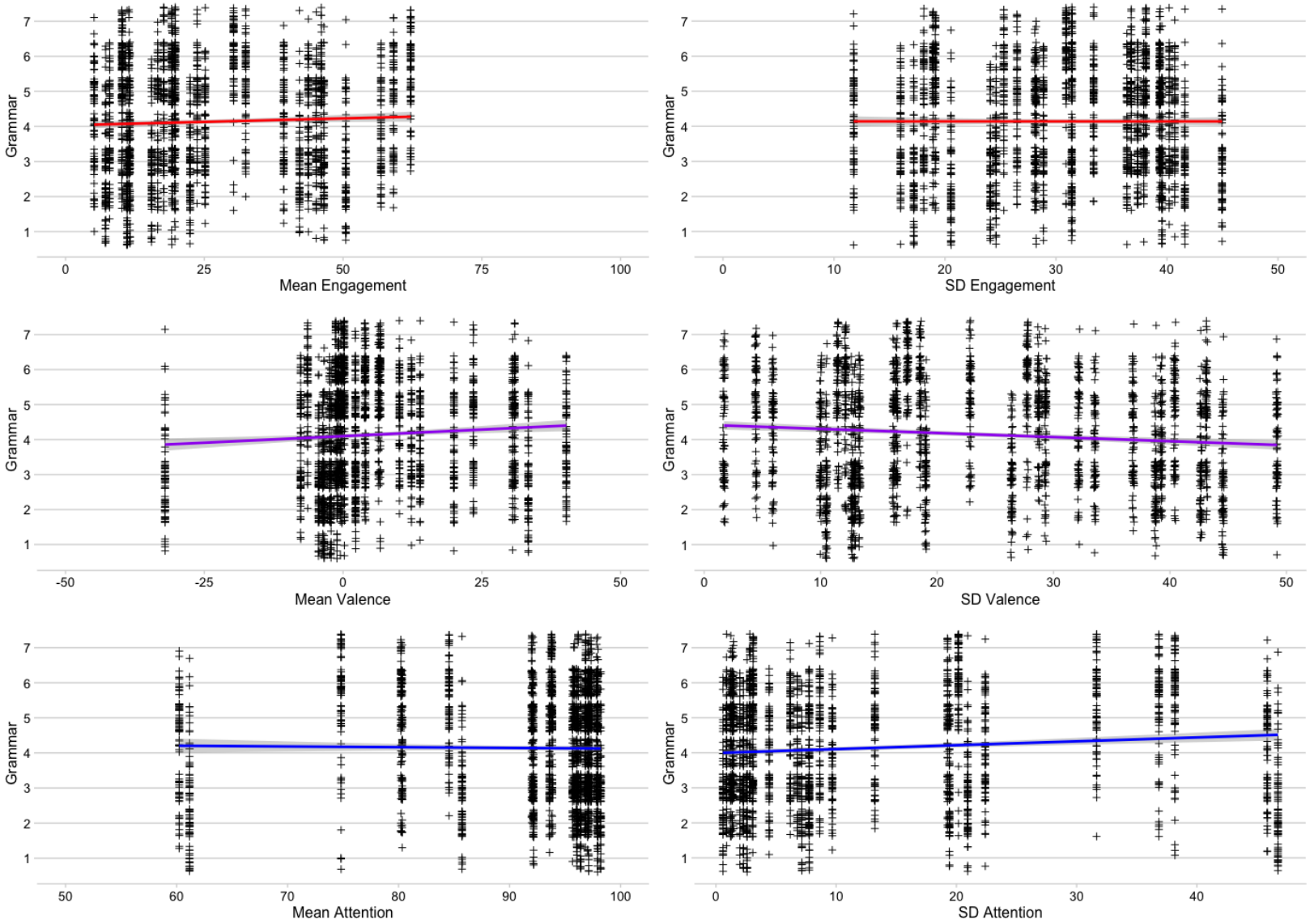
Correlations. Table 6.15 and Figure 6.16 again show similar trends between grammar and the behavioral indices. Here mean valence had an even weaker correlation with grammar (.07, rather than .08 or .12), and a similar correlation between the standard deviations of valence and attention with grammar.

Table 6.15
Polychoric Correlations with Grammar

Variable	Mean	*SD*
Engagement	.04 [0, .09]	0 [-.05, .04]
Valence	.07 [.04, .11]	-.10 [-.14, -.06]
Attention	-.01 [-.06, .04]	.10 [.05, .15]

Figure 6.16

Relationships Between Nonverbal Measures and Grammar



Regressions of main predictors. The results for grammar were analogous to those of fluency and vocabulary. The five-model comparison, shown in Table 6.16, indicated that the best fitting model was the interaction model. The interaction model, presented in Table 6.17, fit significantly better than the other models, $\chi^2(4) = 52.32$, $p < .001$. This model also fit significantly better than the model with random effects removed, $\chi^2(2) = 460$, $p < .001$. As with fluency and grammar, the only significant predictor in this model was base proficiency, which had a sizeable effect on fluency ($\beta = 1.43$, odds ratio = 4.19). This model, as the previous three, only explained minimal variance, Nagelkerke's Pseudo $R^2 = .02$.

Table 6.16
Model Comparisons for Grammar (Means)

Model	AIC	χ^2	*df*	*p*
Null Model	7765			
Model 1	7767	0.38	1	.54
Model 2	7769	0.01	1	.93
Model 3	7771	0.05	1	.82
Interaction model	7726	52.32	4	<.001

Table 6.17
Interactions Between Base Proficiency and Mean Behavioral Indices on Grammar

Coefficients	β	95% CI	*SE*	*z*	*p*	OR	95% CI
Base proficiency	1.43	[0.44, 2.42]	0.51	2.83	.005	4.19	[1.55, 11.30]
Valence	0.03	[-0.01, 0.09]	0.02	1.49	.14	1.04	[0.99, 1.09]
Engagement	-0.02	[-0.06, 0.01]	0.02	-1.24	.21	0.98	[0.94, 1.01]
Attention	0.04	[-0.004, 0.09]	0.02	1.79	.07	1.05	[1.00, 1.10]
Val:Prof	-0.01	[-0.03, -.0002]	0.01	-2.00	.05	0.99	[0.97, 1.00]
Eng:Prof	0.01	[-0.001, 0.02]	0.004	1.68	.09	1.01	[1.00, 1.02]
Att:Prof	-0.01	[-0.02, 0.001]	0.01	-1.62	.10	0.99	[0.98, 1.00]

Random effects

Groups	Variance	SD
Raters	.80	.90
Samples	.31	.56

Note. Adjusted $\alpha = .0125$.

Regression of predictor standard deviations. The five models using standard deviations as predictors, shown in Table 6.18, indicated that the best fitting model was the interaction model, $\chi^2(4) = 55.51$, $p < .001$. This model, shown in Table 6.19, also fit significantly better than the model with random effects removed, $\chi^2(2) = 425$, $p < .001$. This model contrasted with the mean model in that the interaction

of attention with base proficiency was significant, $\beta = 0.01$, $p = .002$, with a very small effect size (odds ratio = 1.01). This model explained minimal variance in the outcome, Nagelkerke's Pseudo $R^2 = .03$.

Table 6.18
Model Comparisons for Grammar (SDs)

Model	AIC	χ^2	*df*	*p*
Null Model	7765			
Model 1	7766	0.85	1	.36
Model 2	7767	0.90	1	.34
Model 3	7767	2.98	1	.08
Interaction model	7719	55.51	4	<.001

Table 6.19
Interactions Between Base Proficiency and Behavioral Index SDs on Grammar

Predictors	β	95% CI	*SE*	*z*	*p*	OR	95% CI
Base proficiency	0.47	[0.02, 0.92]	0.23	2.07	.04	1.61	[1.03, 2.51]
Valence	-0.02	[-0.13, 0.09]	0.06	-0.43	.67	0.98	[0.87, 1.09]
Attention	-0.04	[-0.07, -0.01]	0.02	-2.40	.02	0.96	[0.93, 0.99]
Engagement	0.03	[-0.12, 0.19]	0.08	0.43	.67	1.03	[0.89, 1.21]
Val:Prof	-0.01	[-0.03, 0.02]	0.01	-0.46	.65	1.00	[0.97, 1.02]
Att:Prof	0.01	[0.004, 0.02]	0.004	3.09	.002	1.01	[1.004, 1.02]
Eng:Prof	0.01	[-0.02, 0.03]	0.01	0.38	.71	1.01	[0.98, 1.03]

Random effects

Groups	Variance	SD
Raters	0.80	0.89
Samples	0.23	0.48

Note. Adjusted $\alpha = .0125$.

As in the previous analyses, I calculated the differences in the coefficients and odds ratios between the proficiency groups by each grammar score, presented in Table 6.20. The comparisons were analogous for grammar, with no difference in effect for a score of 4. Figure 6.17 shows the probabilities of a particular score assignment according to the variance in attention, which was effectively the same as with fluency and vocabulary. A higher variance in attention increased a less proficient individual's chance of receiving a 1 or 2 on vocabulary, while lowering their chances of receiving a 5, 6, or 7.

Table 6.20

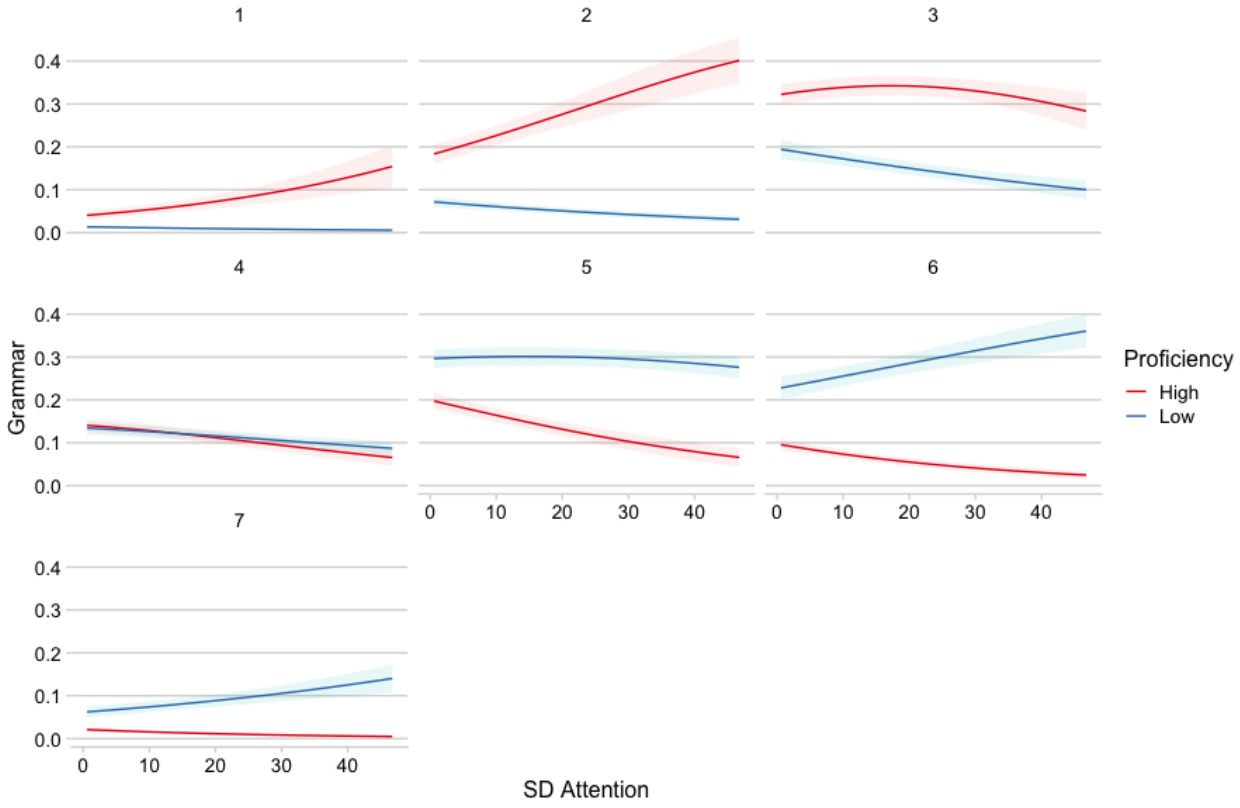
Post hoc Comparisons for Grammar with SD Attention

Score level	*β*	95% CI	*SE*	*z*	*p*	OR	95% CI
1	-0.06	[-0.04, -0.07]	0.01	-8.64	<.001	0.94	[0.93, 0.96]
2	-0.18	[-0.16, -0.20]	0.01	-17.19	<.001	0.83	[0.82, 0.85]
3	-0.15	[-0.13, -0.17]	0.01	-14.57	<.001	0.86	[0.85, 0.88]
4	0.002	[-0.01, -0.01]	0.004	0.37	.70	1.00	[0.99, 1.01]
5	0.13	[0.14, 0.11]	0.01	15.43	<.001	1.14	[1.12, 1.16]
6	0.19	[0.21, 0.17]	0,01	18.14	<.001	1.20	[1.18, 1.23]
7	0.07	[0.08, 0.06]	0.01	11.28	<.001	1.07	[1.06, 1.08]

Note. Adjusted α = .00714 for 7 comparisons

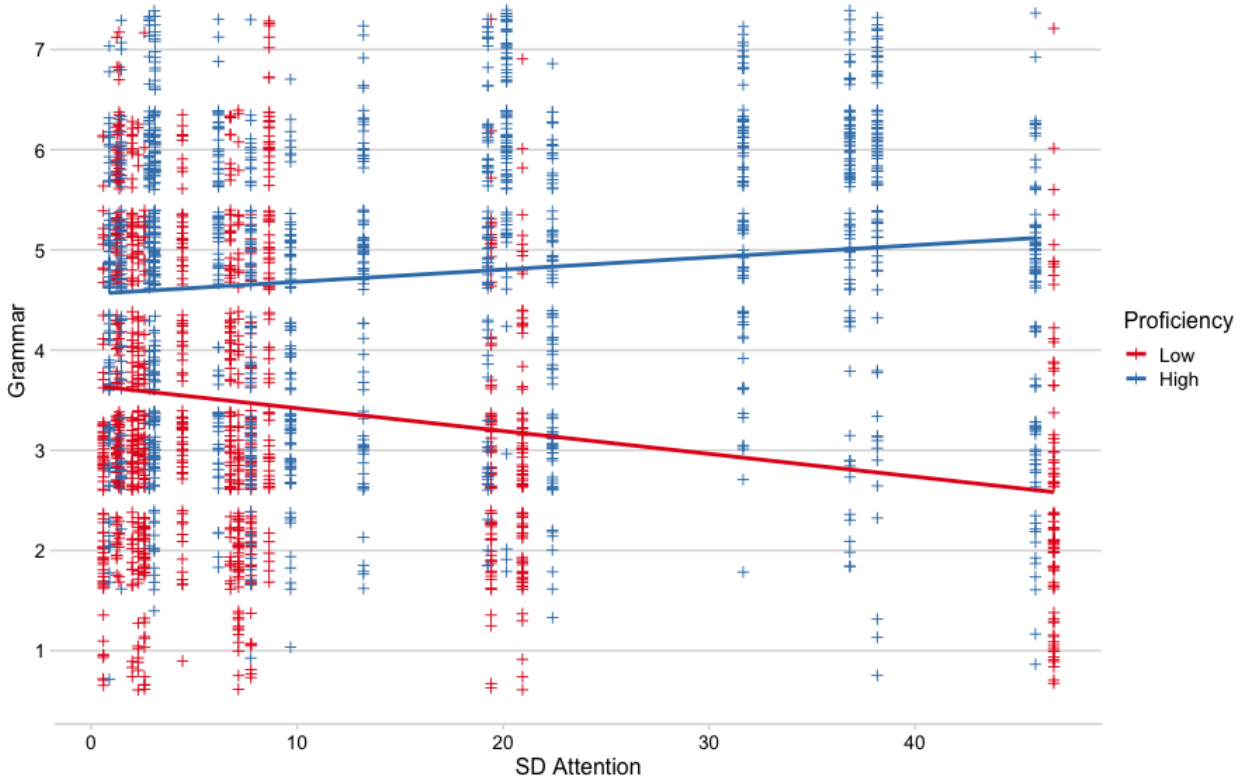
Figure 6.17

Probability of Grammar Score Given SD Attention



I illustrate the effects of this interaction on the dataset as a whole, pictured in Figure 6.18. As with fluency and vocabulary, it can be seen that more varied attention related to differential outcomes for the two proficiency groups, though the impact appears somewhat smaller for grammar.

Figure 6.18
Visualization of Impact of SD Attention on Grammar



Comprehensibility

Correlations. Table 6.21 and Figure 6.19 illustrate the relationships between the iMotion means and standard deviations with comprehensibility. Here trends were slightly different from fluency, vocabulary, and grammar. In addition to a stronger correlation with mean valence (.18, rather than .15, .12, or .07), mean engagement was also a significant positive correlate at .11. This suggests a positive relationship between overall expressiveness and positivity with comprehensibility. Correlations with standard deviations also diverged from fluency, vocabulary, and grammar. Here, valence was not a significant correlate, but engagement and attention were, both correlating positively at .06 and .11, respectively. These findings indicate a positive but small relationship between more varied overall expressiveness and attention with comprehensibility.

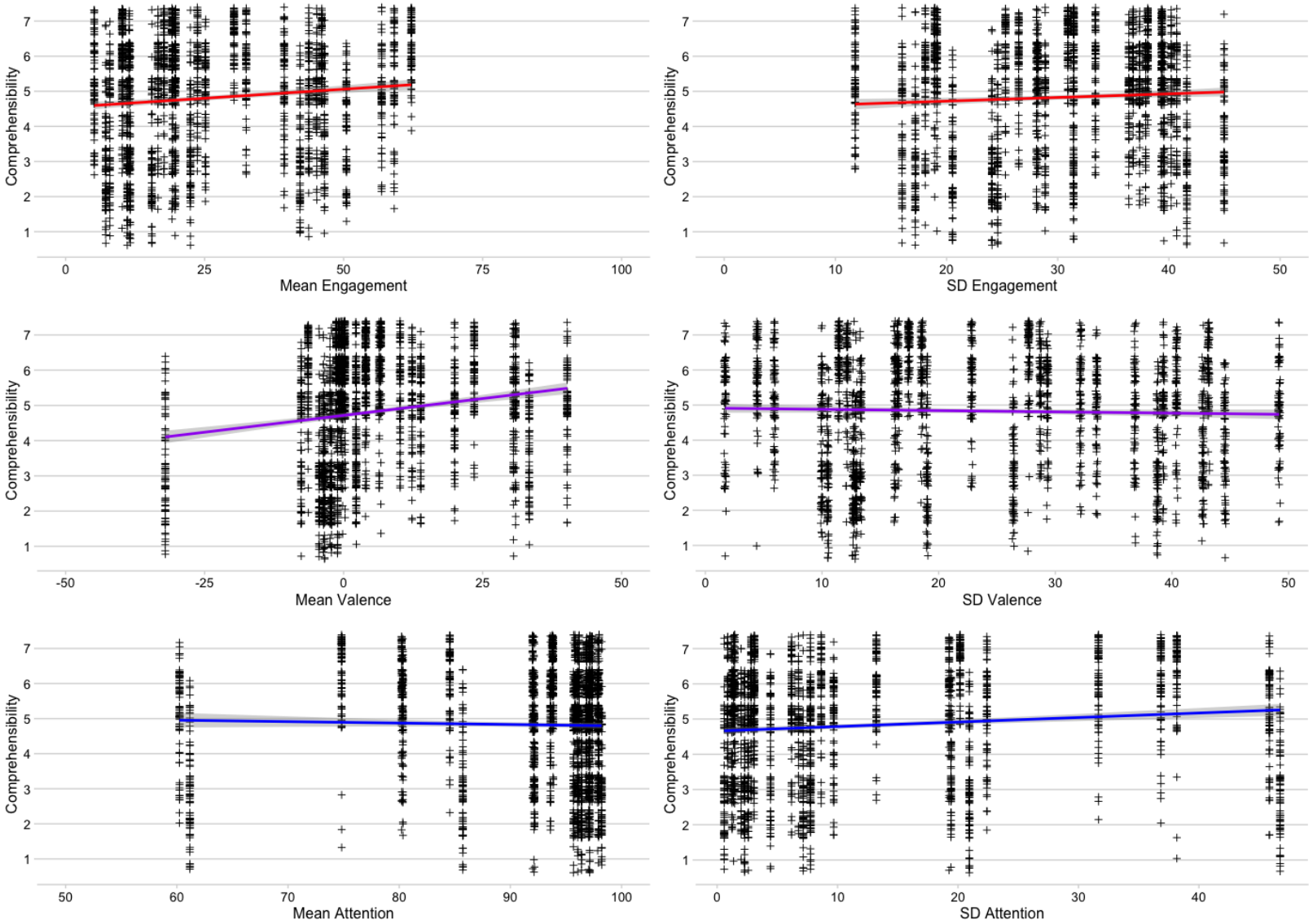
Table 6.21

Polychoric Correlations with Comprehensibility

Variable	Mean	*SD*
Engagement	.11 [.08, .15]	.06 [.02, .10]
Valence	.18 [.14, .21]	-.03 [-.07, .01]
Attention	-.03 [-.08, .02]	.11 [.07, .15]

Figure 6.19

Relationships Between Nonverbal Measures and Comprehensibility



Regressions of main predictors. Factors were entered in the model based on the correlations with vocabulary listed in Table 6.9. The results were nearly identical for fluency. The five-model comparison, shown in Table 6.10, indicated that the best fitting model was the interaction model. The interaction model, presented in Table 6.11, fit significantly better than the other models, $\chi^2(4) = 48.00$, $p < .001$. This model also fit significantly better than the model with random effects removed, χ^2 (2) = 697.42, $p < .001$. This model of mean indices varied substantially from the previous three models. In this model, the interaction between mean valence and base proficiency was significant, $\beta = -0.02$, odds ratio = 0.98, which is a small effect size. The main effects did not reach significance. Similar to the other models, this model explained minimal variance in the score outcome, Nagelkerke's Pseudo $R^2 = .02$.

Table 6.22
Model Comparisons for Comprehensibility (Means)

Model	AIC	χ^2	*df*	*p*
Null Model	7246.90			
Model 1	7247.50	1.41	1	.23
Model 2	7249.40	0.05	1	.82
Model 3	7251.00	0.36	1	.55
Interaction model	7211.00	48.00	4	<.001

Table 6.23
Interactions Between Base Proficiency and Mean Behavioral Indices on Comprehensibility

Coefficients	β	95% CI	*SE*	*z*	*p*	OR	95% CI
Base proficiency	1.48	[0.21, 2.75]	0.65	2.28	.02	4.38	[1.23, 15.59]
Valence	0.08	[0.01, 0.14]	0.03	2.42	.02	1.08	[1.01, 1.15]
Engagement	-0.01	[-0.06, 0.03]	0.02	-0.56	.57	0.99	[0.94, 1.03]
Attention	0.03	[-0.03, 0.09]	0.03	0.87	.38	1.03	[0.97, 1.09]
Val:Prof	-0.02	[-0.04, -0.01]	0.01	-2.57	.01	0.98	[0.92, 0.99]
Eng:Prof	0.01	[-0.004, 0.02]	0.01	1.18	.24	1.01	[1.00, 1.02]
Att:Prof	-0.01	[-0.02, 0.01]	0.01	-1.09	.28	0.99	[0.98, 1.01]

Random effects

Groups	Variance	SD
Raters	1.17	1.08
Samples	0.53	0.73

Note. Adjusted $\alpha = .0125$.

I calculated the differences in the coefficients and odds ratios between the proficiency groups by each comprehensibility score, shown in Table 6.24. The direction of impact shifted much higher for

comprehensibility, between a 5 and 6. Figure 6.20 shows the probabilities of a particular score assignment according to the mean valence value. Here trends were quite different. An increase in mean valence, or the overall positivity of a person's expressions, resulted in lower probabilities of an assignment of 1–3 on comprehensibility and a higher probability of receiving a 5–7. For more proficient speakers, higher valence corresponded with a greater probability of score assignments between 1–5, and surprisingly a lower probability of scoring a 7.

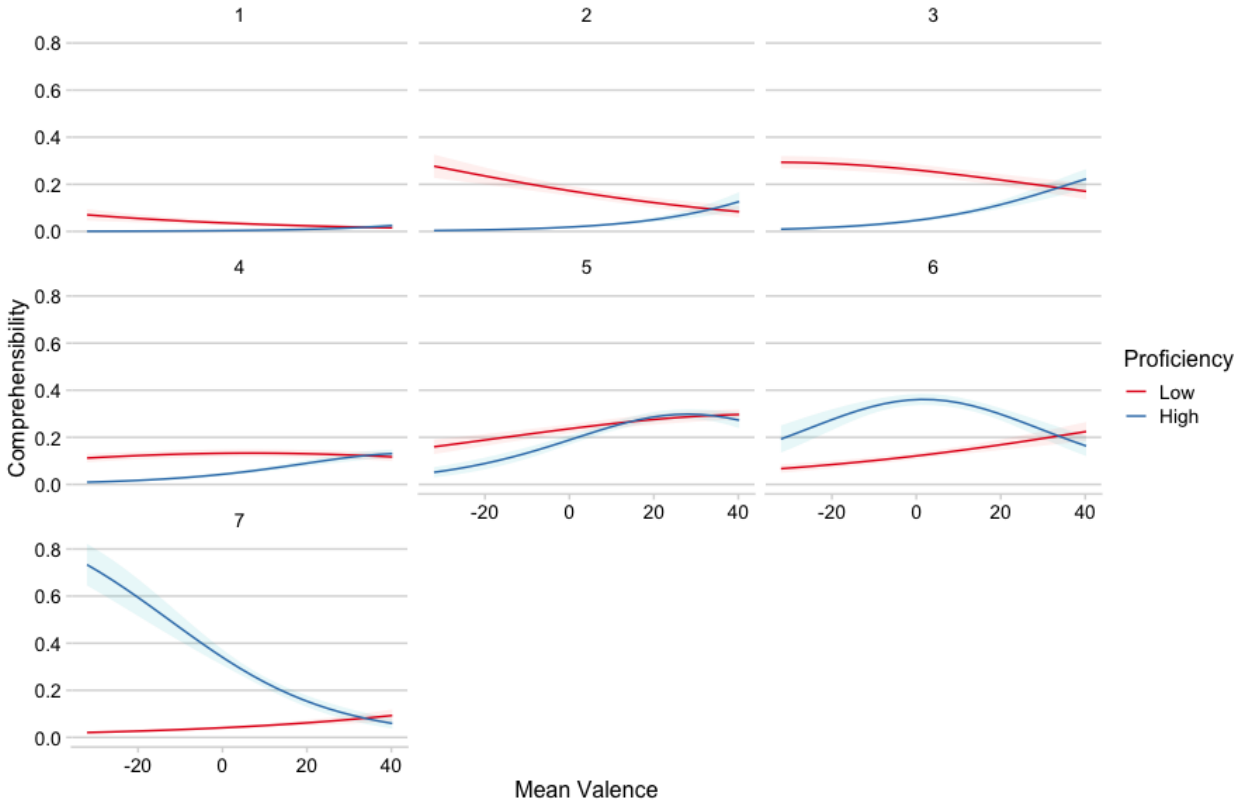
Table 6.24

Post hoc Comparisons for Comprehensibility with Mean Valence

Score level	β	95% CI	*SE*	*z*	*p*	OR	95% CI
1	-0.03	[-0.02, -0.04]	0.004	-6.90	<.001	0.97	[0.96 0.98]
2	-0.13	[-0.12, -0.15]	0.01	-14.64	<.001	0.88	[0.86, 0.89]
3	-0.17	[-0.15, -0.19]	0.01	-17.71	<.001	0.84	[0.82, 0.86]
4	-0.07	[-0.06, -0.08]	0.01	-12.68	<.001	0.93	[0.92, 0.94]
5	-0.03	[-0.01, -0.04]	0.01	-3.30	<.001	0.97	[0.96, 0.99]
6	0.19	[0.21, 0.17]	0,01	18.17	<.001	1.21	[1.19, 1.24]
7	0.24	[0.26, 0.21]	0.01	20.02	<.001	1.27	[1.24, 1.30]

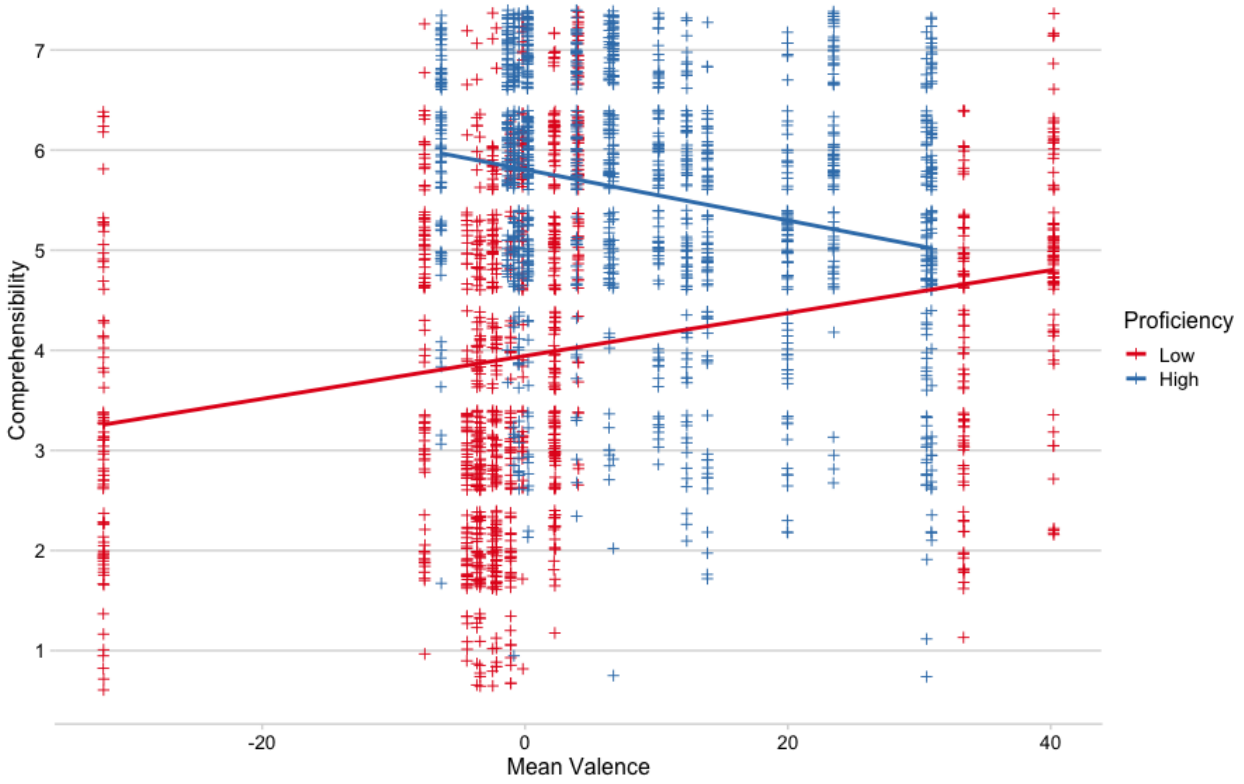
Note. Adjusted α = .00714 for 7 comparisons

Figure 6.20
Probability of Comprehensibility Score Given Mean Valence



The effects of this interaction on the dataset as a whole are illustrated in Figure 6.21. The differential effects are much more apparent here, with a benefit of positive valence corresponding to greater comprehensibility scores in less proficient speakers, and lower comprehensibility scores with higher valence in the more proficient group. It must be said, however, that the range of mean valence was much more restricted in the more proficient group.

Figure 6.21
Visualization of Impact of Mean Valence on Comprehensibility



Regression of predictor standard deviations. Predictor standard deviations were also entered in a secondary model based on the absolute value of correlations with comprehensibility listed in Table 6.21, with the order being attention, engagement, and then valence. The five models, shown in Table 6.25, indicated that the best fitting model was the interaction model, $\chi^2(4) = 46.25$, $p < .001$. This model, shown in Table 6.26, also fit significantly better than the model with random effects removed, $\chi^2(2) = 692.65$, $p < .001$. As opposed to previous models of behavioral standard deviations, none of the predictors in this model were significant. This model explained minimal variance in the outcome, Nagelkerke's Pseudo $R^2 = .02$.

Table 6.25
Model Comparisons for Comprehensibility (SDs)

Model	AIC	χ^2	*df*	*p*
Null Model	7246.90			
Model 1	7247.60	1.25	1	.26
Model 2	7249.30	0.27	1	.60
Model 3	7248.80	2.52	1	.11
Interaction model	7210.60	46.25	4	<.001

Table 6.26
Interactions Between Base Proficiency and Behavioral Index SDs on Comprehensibility

Predictors	β	95% CI	*SE*	*z*	*p*	OR	95% CI
Base proficiency	0.63	[-0.02, 1.27]	0.33	1.90	.06	1.87	[0.98, 3.58]
Attention	-0.03	[-0.08, 0.02]	0.02	-1.27	.20	0.97	[0.93, 1.02]
Engagement	0.02	[-0.20, 0.24]	0.11	0.17	.86	1.02	[0.82, 1.27]
Valence	0.005	[-0.15, 0.16]	0.08	0.06	.95	1.00	[0.86, 1.18]
Att:Prof	0.01	[<.001, 0.02]	0.01	1.98	.05	1.01	[1.00, 1.02]
Eng:Prof	0.01	[-0.03, 0.05]	0.02	0.44	.66	1.01	[0.97, 1.05]
Val:Prof	-0.001	[-0.04, 0.02]	0.02	-0.62	.54	0.99	[0.96, 1.02]

Random effects

Groups	Variance	SD
Raters	1.17	1.08
Samples	0.53	0.73

Note. Adjusted α = .0125.

Summary

The analysis in this chapter has utilized cutting edge pattern recognition software—iMotions, running Affectiva—to analyze human behavior in an online speaking test. This study is only the second of its kind, at the time of writing this paper, to use automated facial expression analysis in a study of nonverbal behavior in applied linguistics (after Chong & Aryadoust, 2023). The software extracted behavioral indices of engagement, or overall expressiveness, valence, and measures of attention from 30 video samples. These objectively derived indices, as opposed to subjective ratings of affect in Chapter 5, were used as predictors in models to determine their overall impact on language proficiency scores. These indices were calculated as mean overall values of the behavior over the course of the speaking test, which were hypothesized in the preregistration of this study to impact language proficiency scores. As an additional exploratory analysis, I

also ran models with the standard deviations of the predictors as a measure of how much each behavior varied during the test sample.

The study showed that engagement (as measured by facial expressiveness), valence, and attention do not have an impact on language scores without base proficiency taken into account. Base proficiency, which consisted of the scaled IELTS scores, interacted with these measures to produce differential outcomes. Greater positive valence, or the emotional expressiveness of a test taker, related to less proficient speakers being perceived as more comprehensible, while having a negligible relationship with more proficient speakers. On the other hand, more varied attentional focus, which was measured through head and gaze turns away from the examiner/camera, corresponded with lower fluency, vocabulary, and grammar scores in less proficient speakers, and somewhat higher scores in more proficient speakers.

In the next chapter, I turn to the raters themselves. While the first two studies have extrapolated about the effects of various metrics on proficiency outcomes, Chapter 7 will consider what the raters were thinking as they assigned scores to the speaking test samples. This source of data is critical to a fuller understanding of the phenomena at play, as rater reports may offer insight that triangulates and thus supports findings from the quantitative analyses.

CHAPTER 7: NONVERBAL BEHAVIOR AND RATER COGNITION

The past two chapters have explored trends within the variables observed by the raters and between the externally measured iMotions data and the languages scores. While the quantitative analysis has revealed interesting trends, developing a greater understanding requires introspection in the form of rater reports. In this chapter, I describe the results of a stimulated verbal recall I conducted in order to triangulate the findings of the quantitative analyses. I conducted recalls with 20 of the undergraduate, untrained participants from the main study. These raters provided reasons for their judgements on 10 files. These files were selected using Rasch analysis, where I determined which files had changed their relative ordering with the original test scores the most. I hypothesized that these files, where the lay raters disagreed the most with the trained raters, would be impacted by criteria outside of the language test construct to a greater degree. The research questions guiding this chapter are as follows:

RQ3.1: Which nonverbal behaviors are most salient and informative to raters when scoring?

RQ3.2: How do raters understand language proficiency in light of nonverbal behavior?

In this chapter, I will first describe general trends in the stimulated recall data relating to comments raters made as a whole. I will then provide descriptive data on the frequency of occurrence of nonverbal behaviors that the raters observed. Following this, I will investigate deeper patterns in the dataset relating to how raters used nonverbal behavior to formulate judgements related to language proficiency.

General rating trends

The general aim of the stimulated verbal recall in this study was to elicit memories from participants about their thought processes during the online sample rating. At the same time, the clandestine interest of the recall was to understand how they used nonverbal behavior during their judgements, of which raters were not aware. Because the rating scale used in this study contained elements relating to language followed by affect, post-hoc evidence of the validity of the procedure should show that raters focused primarily on language but also on affect. While raters focusing primarily on nonverbal behavior may not necessarily be evidence of their awareness of the questions guiding the study, any such cases should be inspected carefully. In this section, I provide validation evidence of the recall while highlighting overarching trends in the raters'

comments.

The raters produced comments that resulted in 4,184 coding decisions. The total count of each coding category, presented in Table 7.1, provides evidence of where raters directed their attention. Raters commented the most on language features (38%), followed by affect (31%), aspects of the test interaction (20%), and finally nonverbal behavior (11%). These results indicate that the participants indeed provided the most commentary on elements related to the rating scales, as intended by the procedure and indicated in the instructions. However, raters also made a sizable number of comments about the test interaction, in particular concerning the content of the test takers' discourse (e.g., ideas mentioned, the topic, truthfulness, amount and breadth, etc). For example, when discussing sample 21, rater 2 makes comments about the content of speech and overall comprehensibility:

> *She seems like she's having a really bad time trying to come up with an answer to his question. She says that people can do the same thing. That's it's pretty vague and nonspecific. It's hard to tell what she may mean by what she's saying. So some ideas are there, but they're not very descriptive. So they can't be easily understood.* (Content, comprehensibility)

Nonverbal behavior occupied the smallest number of coding comments. This suggests that raters were not, as whole, previously exposed to or aware of the research questions. While 11% is a relatively small percentage in comparison, it aligns with findings from Sato and McNamara (2019) where lay raters offered recall on their rating decisions of communicative ability. In that study, comments on CET-SET language features comprised 36.7%, content 15%, and nonverbal behavior 9.4%. Affect, coded as composure/attitude, only made up 5.7% of the comments, but this discrepancy between these two findings is certainly due to the inclusion of affect on the rating scales. Similarly, in May (2011), (trained) raters' comments on features of nonverbal behavior relevant to interactional competence reached a similar figure of 12%. Based on these results, the findings in this study align extremely well with previous work, which provides evidence of the veridicality of the nonverbal observations within the stimulated recall data.

Table 7.1
Code Counts for Stimulated Recall Data

Code	Count	%*	Code	Count	%*
Affect	**1,278**	**31**	**Nonverbal Behavior**	**477**	**11**
Anxiety	215	17	Gaze	115	25
Confidence	183	14	Mouth	96	21
Happiness	126	10	Paralinguistics	75	16
Competence	113	9	Face (General)	53	11
Engagement	109	9	Posture	52	11
Expressiveness	121	9	Gesture	35	7
Warmth	96	8	Head	30	6
Attentiveness	92	7	Eyebrows	11	2
Attitude	93	7	Body (General)	10	2
Interactiveness	66	5			
Desire to Communicate	37	3			
Humor	27	2			
Language	**1,576**	**38**	**Test Interaction**	**853**	**20**
Fluency	308	19	Content	368	43
Comprehension	244	15	Thinking	150	18
Comprehensibility	228	14	Turn-taking	99	12
Vocabulary	294	14	Examiner	67	8
Grammar	224	13	Repair	65	8
Overall Ability	145	9	Relevance-Contingence	57	7
Pronunciation	101	5	Active Listening	35	4
Organization	32	2	Visual Artifacts	12	1

**Note*. Percentages in bold are percentages of total 4,184 comments. Percentages for each subcode are percentages of the total of the grouping code.

In this dataset, all raters made comments on these four broad categories, albeit to different degrees. Table 7.2 shows the distribution of these comments as percentages of their total number of comments. Raters made an average of 124.6 coded comments each (*SD* = 39.01), ranging from as few as 87 to as many as 224. All but four raters focused on language the most, ranging from 29–66% of their comments. Raters 03, 08, 14, and 20 deviated from these trends somewhat, focusing more on test interaction (rater 03; 36%) or affect (raters 08, 14, and 20; 31–34%). There were no raters that focused the most on nonverbal behavior, and there were no apparent patterns attributable to gender in how raters commented on the speech samples. Raters were idiosyncratic, however, in how much they commented on nonverbal behavior. Comments ranged from as few as 5% to as many as 24%. Overall, despite these differences, raters appeared to adopt a similar focus across their stimulated recalls.

Table 7.2
Percentages of Coded Comments by Raters

SVR Rater	Raw total	Language	Affect	Test Interaction	Nonverbal Behavior
01	101	**66**	5	24	5
02	95	**39**	18	37	6
03	84	27	29	**36**	8
04	88	**36**	18	30	16
05	142	**42**	20	29	10
06	133	**39**	31	20	10
07	224	**31**	29	24	17
08	222	30	**31**	22	17
09	87	**44**	21	29	7
10	107	**31**	27	18	24
11	129	**37**	31	21	11
12	110	**34**	**34**	20	13
13	120	**38**	23	34	5
14	136	28	**32**	18	22
15	97	**38**	22	28	12
16	152	**33**	26	24	18
17	93	**33**	30	18	18
18	115	**43**	23	23	10
19	131	**29**	27	26	18
20	125	32	**34**	22	12

**Note*: Bold indicates the highest of the four coding categories per rater.

Salience of nonverbal behavior

As documented in the previous section, all raters indeed found nonverbal behavior salient to their rating processes. The features the raters noticed, however, differed substantially. Table 7.3 lists the nonverbal behaviors raters commented on throughout the stimulated recalls, categorized by area of the body and type of behavior. In each set, the percentage of the categorical code is that of the total number of behaviors observed, 477; the percentages of the subcodes of each specific behavior pertain only to that grouping. For example, comments on gaze made up 25% of the 477 comments on nonverbal behavior, while shifting gaze made up 42% of the 115 comments on gaze. Importantly, this table also includes the extensiveness of the comments, representing the number of raters that commented on this area. Thus, for example, 19 raters made comments on some aspect of gaze, while 16 raters commented specifically on

shifting gaze.

Table 7.3
Coding Counts of Nonverbal Behavior

Code	Count	%	Ext	Code	Count	%	Ext
Gaze	**115**	**24**	**19**	**Mouth**	**96**	**20**	**20**
Shifting	50	42	16	(Genuine) smile	64	63	20
Mutual	31	26	10	Lack of smile	12	12	4
Averted	28	24	8	Lip movements	9	9	6
Staring	6	5	4	Nervous smile	6	6	4
Eyes grow wide	2	2	2	Swallowing	6	6	4
Blinking	1	1	1	Mouth barely open	2	2	2
Unfocused	1	1	1	Frowning	2	2	1
Paralinguistics	**75**	**16**	**19**	**General face behaviors**	**53**	**11**	**17**
Production speed	88	49	19				
Filled pauses	35	19	14	**Posture**	**52**	**11**	**17**
Laughing	27	15	15	Rocking/Shaking	14	24	10
Tone-prosody	23	13	8	Leaning forward	12	20	7
Audible breathing	5	3	2	Moving around	11	19	5
Backchanneling	1	1	1	Rigid/Straight posture	9	15	7
Volume	1	1	1	Leaning back/Slouching	6	10	5
				Adjusting posture	4	7	4
Gesture	**35**	**7**	**13**	Shoulder movements	3	5	2
Self-adaptors	15	43	9				
Representational gestures	9	26	5	**Head**	**30**	**6**	**13**
Lack of hand movement	7	20	4	Nodding	23	74	11
Random movement	4	11	4	Turns	8	26	5
Eyebrows	**11**	**2**	**6**	**General body language**	**10**	**2**	**6**
Movement	6	55	3				
Furrowed	3	27	2				
Raised	2	18	2				

Note: "Ext" refers to the extensiveness of appearance: the number of raters that mentioned this feature

Table 7.3 shows several patterns relating to what raters discussed when observing test takers. Gaze and eye behaviors were the most common behavior topic in this dataset, at 24% of the comments. Nearly all raters commented on this behavior, likely due to its visual salience in the online space where generally only the face is seen (Batty, 2021). Closely following gaze were mouth behaviors, making up 20% of the comments and being commented on by all 20 raters. This is also anticipated, as the mouth serves to produce language, and individuals often look at the mouth to enhance comprehension (in L1 speakers, Krason et al., 2022; and L2 speakers, Batty, 2021; Hardison, 2018; Hardison & Pennington, 2021; Ockey, 2007; Worster

et al., 2018) and to interpret affect (Coniam, 2001). Paralinguistic features, which I defined quite broadly as sounds or features of speech not explicitly associated with verbalizations, made up 16% of the comments on nonverbal behavior. These were observed by nearly all raters. Similarly, non-specific references to the face were also quite extensive, being observed by 17 raters, making up 11% of the comments. Posture (11% of comments, 17 raters), gesture (7% of comments, 13 raters), and head observations (6% of comments, 13 raters) were also observed by more than half of the raters. This is somewhat surprising for gesture and posture, as each are sometimes harder to see in the online format. Finally, eyebrow movements and non-specific references to body language each comprised 2% of the dataset and were mentioned by fewer than half of the raters. Thus, the most salient behaviors related to either the eyes, the mouth, or sounds from the mouth, making up over half of all observations. Comments about inexpressiveness or the lack of particular behaviors were present, but not extremely common.

In this dataset, however, comments about nonverbal behavior were almost always accompanied by evaluations related to affect. There is precedent in this, as Sato and McNamara (2019) noted that in their study "[t]he speakers' [nonverbal behavior] and composure/attitude were intertwined, since confidence was judged primarily through observed [nonverbal behavior]" (p. 908). It is then worthwhile when describing behaviors in the dataset to highlight where each behavior intersected with judgements of affect. These intersections are available in Appendix J. Table J.1 displays intersection percentages across behaviors; that is to say, for each behavior it indicates the percent of intersections for each of the 11 affect judgements on that particular behavior. For example, 33% anxiety with body language in this table indicates that 33% of comments, out of 19 total, about body language and affect indicated a comment about anxiety or relaxedness/ease. The only other large intersection in this row is with confidence at 28%. Table J.2 considers the same data but within each category; that is to say, it indicates the comments about behaviors that coincided the most with each individual affect judgement. In anxiety, for example, shifting gaze coincided with anxiety in 13% of all behavioral comments, out of a total of 217 intersections. It is important to note that a coincidence with anxiety may indicate both valences of anxiety or at ease, and likewise for all other affect judgements. Thus, it is critical to consider the comments themselves when extrapolating

about whether evaluations were positive or negative.

Gaze

Gaze was the most frequently mentioned behavior in this dataset, making up a quarter of the comments and being observed by 19 of the 20 raters. There was an average of 5.75 comments about gaze per rater in the pool of 19, with a standard deviation of 5.23. Stimulated recall participant-raters observed seven types of gaze behaviors, but three made up the bulk of all comments: shifting gaze, mutual gaze, and averted gaze. Less frequently mentioned were staring, eyes growing wide, blinking, and unfocused gaze. Shifting gaze, described as looking around or eyes darting all over, was the most salient gaze phenomenon, making up 42% of the gaze comments and being observed by 16 raters. For example, rater 17 remembered aspects of sample 16's performance almost immediately:

> *Um, I remember him kind of looking around a lot. And I thought of that as like, you know, he's trying to remember things, you know. And he was stuttering a little bit. So I remember thinking he was anxious.* (Shifting gaze, anxiety, fluency, thinking).

Figure 7.1
Annotation Density Plot for Sample 16

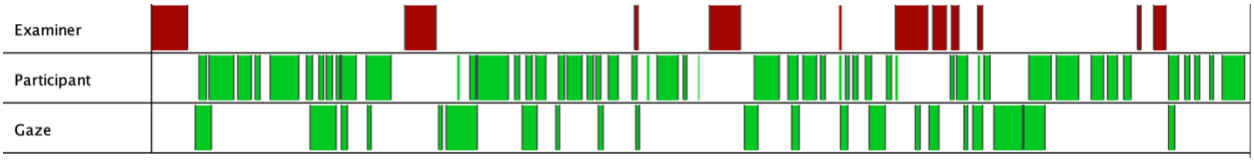


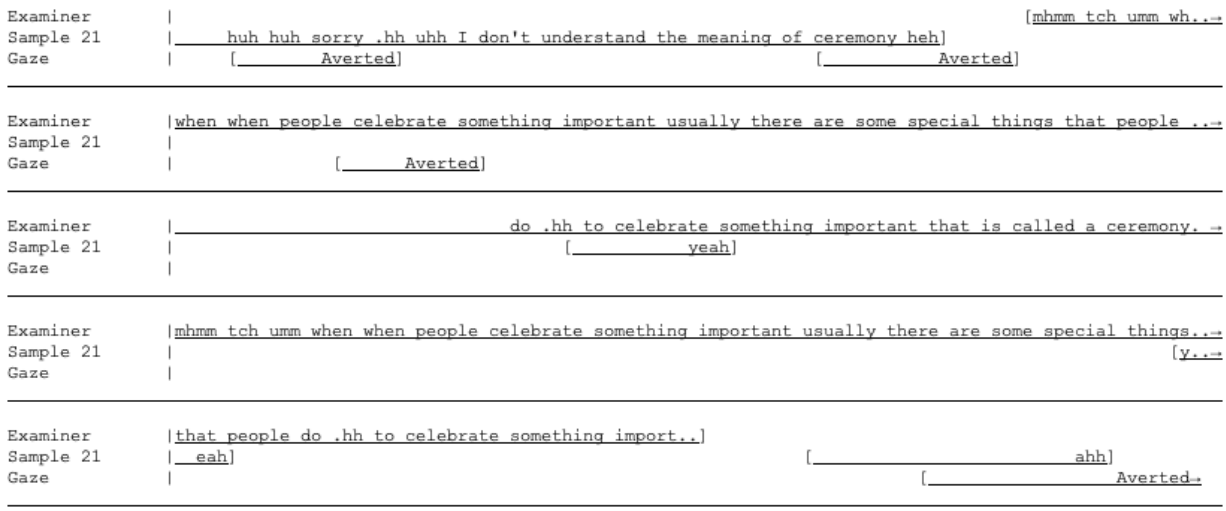
Figure 7.1 shows the annotation density plot for sample 16. The tiers labeled Examiner and Participant show units of sustained speech, while the Gaze tier shows all the moments the test taker was looking away. The plot shows that the test taker frequently shifted between looking away and looking at the examiner. The rater found the test takers' shifting gaze to signal that the test taker was searching for information, and its coincidence with the fluency pattern of stuttering (likely referring to false starts or repair) led to an overall negative view of the test taker. In fact, shifting gaze was almost always considered negatively, as it coincided with anxiety 36% of the times both co-occurred, and confidence (lack of, in this case) 32%. This type of eye movement, however, differed markedly from mutual or averted gaze.

Mutual gaze, that of maintaining eye contact with the unseen examiner in the video, was also

mentioned frequently at 26% of the gaze comments. It was generally seen positively, often coinciding with mentions of engagement (52%), attention (35%), and confidence (19%). For example, when describing sample 21, rater 16 said:

> *Okay. So here, just like in the first 10… like seconds, or whatever this is, she's like clearly focused in on what he's saying. Like her eye contact is dead on the screen. And she's also saying like, "yes", like after he says something. So it clearly shows that she is an active listener, and she is like engaging herself in the conversation.* (Mutual gaze, engagement, comprehension, turn-taking, active listening)

Figure 7.2
Sample 21's Gaze Patterns During Repair Sequence



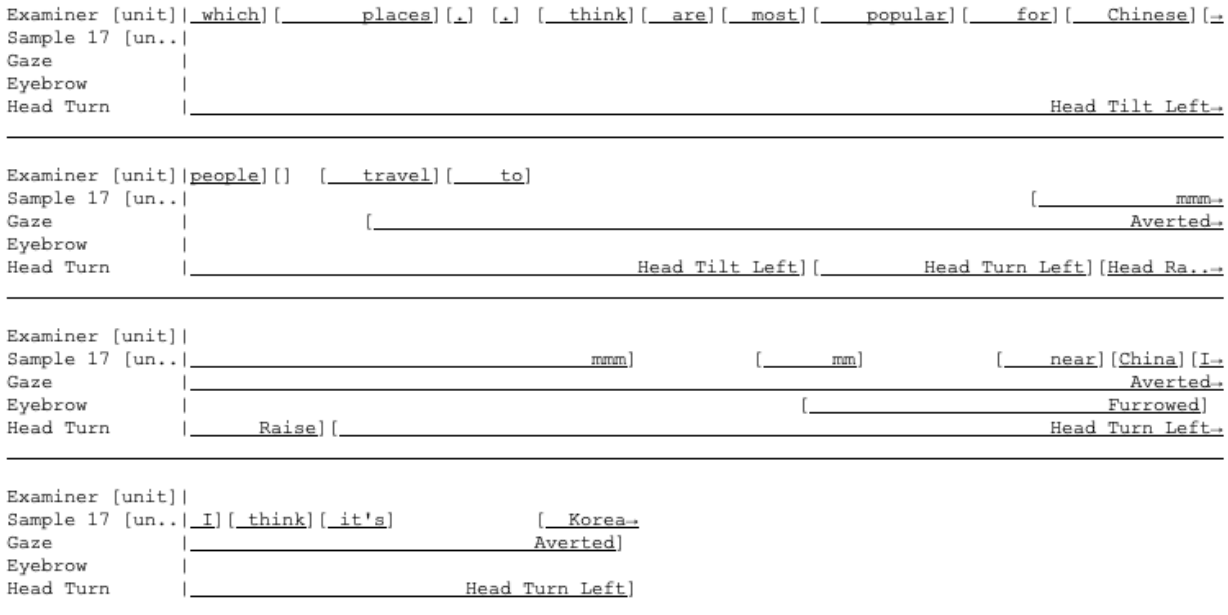
Here, even though this test taker struggled throughout the sample to understand test questions, the rater's view of the test taker was generally positive because the test taker engaged with the examiner by maintaining eye gaze and actively backchanneling. This is visible in Figure 7.2, where the test taker averted her gaze three times while asking for a clarification and at the beginning of the examiner's repair sequence. This was followed by 16 seconds of unbroken mutual gaze (mutual gaze is identified by unannotated gaze segments) accompanied by verbal backchanneling ("yeah"), which was only averted once comprehension was resolved ("ahh"). In other words, mutual gaze here along with the test taker's verbal interactional moves contributed to the rater's evaluation of the test taker's listening skills and engagement in the test, despite

ongoing problems with comprehension.

Averted gaze, maintaining gaze away from the examiner but without shifting the gaze location, made up the third most frequently mentioned gaze behavior at 24%. Depending on the context, however, it was viewed differently. When seen as an act of processing the test content and preparing an answer, it often coincided with engagement, being mentioned as a sign of thinking about the question. For example, rater 4 described sample 17 as:

> *Really engaged, being close. She's heavily thinking about the question, and she understands it usually pretty quickly. Like it doesn't take her too long to respond. Even though she's looking off to the side, she's still like, she's really engaged and thinking about the question. And like I've heard... hers felt more natural when she was talking, like she really understood him.* (Averted gaze, engagement, thinking, comprehension, turn-taking, speed, posture, content)

Figure 7.3
Sample 17's Averted Gaze Upon Comprehension



The test taker's gaze patterns the rater mentions are visible in Figure 7.3. Before the question ended, the test taker showed her comprehension of the question by averting her gaze, followed by two filled pauses before she began her response. During this entire sequence, she maintained averted gaze, thus signaling she

was thinking and preparing her response. She reestablished mutual gaze when she arrived at the direct answer to the question ("Korea"), indicating to the examiner the importance of this key word. In this example, however, it is likely the co-occurrence of the head turns and the furrowed eyebrows along with averted gaze that led the rater to conclude that the test taker was listening and thinking. The fact that the response was natural and contingent on the question likely led to an overall positive evaluation.

However, a positive evaluation of averted gaze was not always the case, as demonstrated by rater 14 when describing sample 21:

> *She seemed very unconfident in even trying to, like, try and answer it. She asked right away. … Right as he started talking, she looked away, and was immediately thinking. And then, as soon as he was done, she said, "Sorry", so she apologized for not knowing something.* (Averted gaze, confidence, turn-taking, comprehension, repair, speed)

Here, averted gaze was also tied closely to thinking, but because the sequence was followed by a failure to comprehend the question and other-initiated repair, the rater overall considered this sequence as relating to a lack of confidence. Averted gaze has also been reported in the L2 literature as relating to anxiety (Gregersen, 2005). Thus, while averted gaze related to evidence of the same cognitive process (thinking), the evaluations based on it were distinct.

The other gaze behaviors—staring, eyes open wide, blinking, and having unfocused gaze—were all mentioned relatively infrequently, occurring only 10 times in this dataset. Apart from blinking, which was an observation with no value judgement, the other behaviors were seen negatively as a sign of internal struggle to perform during the test. For example, rater 12 commented on sample 16's intense stare, finding it to be so focused on the test question that it came off as somewhat distant:

> *I think this stare, kind of more unhappy, but more, because I think he was so focused and like, so attuned to what he was doing that he kind of like. Yeah, you just get so focused, and you want to get it all out kind of, but a little bit colder.* (Staring, happiness, warmth)

Similarly, sample 18's wide eyed looks prompted rater 14 to remark that:

> *There is where I kind of noticed she wasn't completely at ease, because every time he asked a*

question, she got a little wide eyed. And it seemed like she was worried she wasn't going to know what it was after the second time. (Eyes grow wide, anxiety, comprehension)

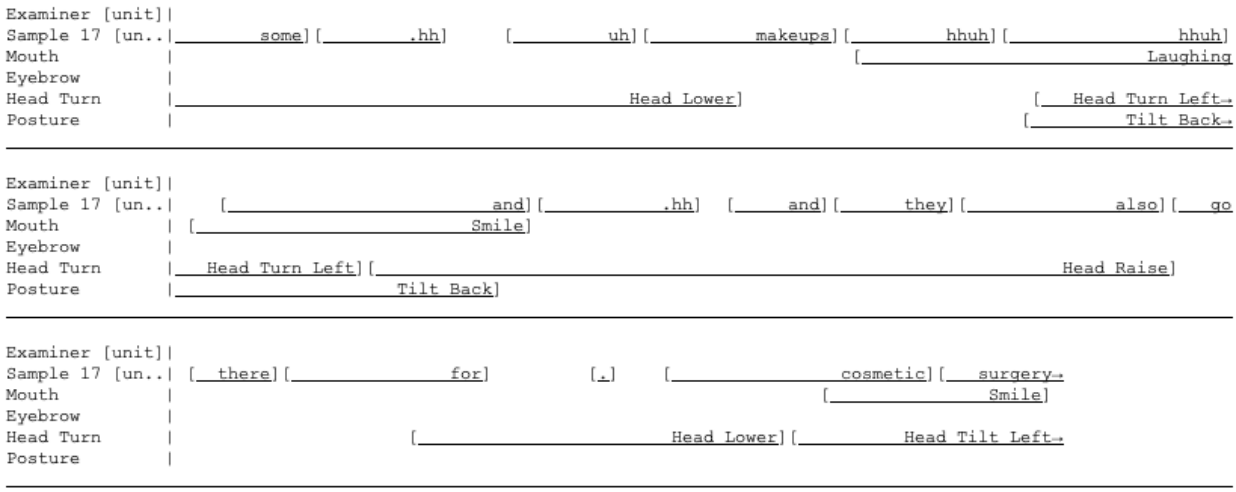
For this rater, the test taker's eyes growing wide was not necessarily a sign of a breakdown in understanding, but rather an affective response to the intensity of listening and the social expectation to respond contingently. In both cases, the lack of these behaviors was also noted as a positive attribute.

Mouth

Mouth movements were the second-most frequently observed category of behavior, making up 20% of the comments. These were observed by all 20 raters (M = 4.80 comments per rater, SD = 4.00). Rater-participants observed seven mouth behaviors, with most comments concerning smiles, a lack of smiling, and lip movements, with fewer comments on nervous smiles, swallowing, a relatively closed mouth, and frowning. The vast majority of these comments, 63%, were about smiles, in particular smiles that were perceived as genuine and not nervous. All raters noted the presence of smiles, making this the most extensively mentioned behavior. Smiling, when not perceived as a nervous smile, was almost always associated with positive evaluations, notably being happy (20%), warm (11%), and having a positive attitude (13%). This is illustrated by rater 13 discussing sample 17:

So I put happy more towards happy just because she seemed like, I don't know, she seemed like she was willing to answer things. And like, she obviously like laughed and like, smiled when she was talking about the makeup in Korea. … And I said, positive attitude, because I don't know, I guess it's like, kind of the same thing is like, the warmth scale is like, if somebody's like smiley and like, outgoing. I'd say that they have a good attitude. (Smiling, happiness, desire to communicate, content, attitude, warmth)

Figure 7.4
Sample 17 Smiling and Laughing



As rater 13 noted here, the test taker began laughing once she brought up the idea about makeup, as seen in Figure 7.4, accompanied by a head turn and leaning back in her chair. This laughing was followed by an episode of smiling, and a smile once again once she mentioned cosmetic surgery. The presence of smiling and laughing, as well as possibly the relaxed posture and head movements, led the rater to perceive this test taker as happy, warm, and with a positive attitude. It is notable that the test taker's desire to communicate, identified as "she was willing to answer things," was also a factor when perceiving this performance as happy. Smiling was also associated with being at ease rather than anxious or nervous, as mentioned by rater 10 on sample 18:

> *So at first, she definitely seemed nervous, but then as it went on, she starts kind of smiling. So she definitely seemed more at ease.* (Smiling, anxiety)

In this case, smiling appeared to be evidence of the exact opposite of anxiety, though raters also detected nervous smiles in the dataset.

Smiling also led some individuals to see performances as being more overall engaged (6%), as was the case of sample 15 when rater 4 remarked:

> *I was, I think I felt really, really good. Because she was smiling the whole time. And as she was talking about her mom and stuff. It felt like she was very engaged in what she was talking about.* (Smiling, content, engagement)

The rater here noted the close association between content (a memory about her mother) and the test taker's reaction by smiling. This led the rater to feel this story was genuine, and that she was engaging with something real that happened in her life. Interestingly, the positive affect from the video appears to have impacted the rater as well when the rater noted "I felt really, really good." Although an infrequent observation in the dataset, this comment may be evidence of emotional contagion (Elfenbein, 2014; Hatfield et al., 1994), with the rater being emotionally influenced by the behaviors witnessed.

The lack of a smile made up the second highest number of mouth comments at 12%, although this was only mentioned by four raters. This behavior, or rather its absence, did not indicate that the test taker was frowning or exhibiting any other mouth behavior. Rather, it was usually mentioned when raters were struggling to understand the test taker's underlying happiness or attitude, such as rater 7 when describing sample 17:

> *I don't think he smiled too much. Yeah, that one was kind of a hard one to gauge at the same time with the happy/unhappy. But you know, there's a lot of things that go into that for sure. … I think I gauged a little bit on their actual responses, because that's a time you could see if their answer is competent and they spell it out and everything. You can see they're happy to be answering that question. But there was a question about, for example, the girl who didn't understand "ceremony" you know, she didn't seem super pleased to be answering….* (Smiling, content, competence, happiness, turn-taking)

In this extract, the rater lacked nonverbal evidence of happiness in the forms of smiling yet was hesitant to determine that the test taker was unhappy based on this evidence alone. The rater resorted to inferring the test taker's affective state through their ability to understand and answer the test question, and the effectiveness of the response. Competence for this rater, in the form of overall ability, was an alternate line of evidence to infer underlying attitude.

Lip movements formed the third highest number of comments from raters, at 9%. These constituted their own range of behaviors, such as lip biting, "dry mouthing," smacking lips, and random lip movements. These were generally associated with anxiety, such as with sample 21, described by rater 14:

> *And then, you, I could see her, she's back to touching her face. And she like bit her lips, making her seem anxious and unconfident. But she is always engaged when he's talking. She's looking at the screen and thinking ahead to what she's going to try and say about what he's talking.* (Lip movements, self-adaptors, anxiety, confidence, engagement, mutual gaze, thinking)

In this case and in others, anxiety and a lack of confidence was perceived through the test taker's lip biting and self-adaptors, but the overall evaluation was balanced by the fact that the individual maintained eye contact and appeared engaged with the test. Thus, though some of these behaviors often led to negative impressions, the overall impression was positive likely because the test taker is actively participating and engaging in the speaking test. Like averted gaze, the impact of behavior was mediated by context.

A smaller number of comments concerned other mouth behaviors which were less frequently mentioned. These behaviors also coincided largely with being anxious and a lack of confidence, but they did not always result in an overall negative impression. For example, when rater 20 described sample 21's nervous smile, he still found her to be overall positive, albeit not necessarily happy:

> *I did rate her a little bit higher on this part, you know, personableness and warmth. But I don't think that she was particularly happy in this experience, though. Despite her smiling, that definitely feels like an anxious smile to me, which again plays into anxiety score.* (Nervous smile, warmth, happiness, anxiety)

Common throughout these comments is that stimulated recall participants were regularly evaluating multiple lines of evidence when understanding the test taker's affective states, often both nonverbal and verbal.

Paralinguistics

Paralinguistic features made up the third most frequently commented phenomenon, making up 16% of the total comments and spread across 19 raters (M = 3.75 comments per rater, SD = 2.45). These features

comprised sounds made from test takers that were not verbal; that is, not composed of words with specific form-meaning relationships. While silence or a reticence to speak was mentioned in the dataset, I did not code this phenomenon because it was inherently not a sound and did not indicate the lack of any specific behavior apart from speech. The most common paralinguistic feature mentioned by the raters was that of production speed, or the rate of articulation of utterances. This feature was commented on 88 times (49%) by 19 raters. It was often mentioned in relationship to fluency or the test taker's overall ability. Rater 1 made this connect quite explicitly when discussing sample 9:

> *It was a pretty complicated question. And so when she answered it quickly, it made me feel a little bit better about her fluency. Because I guess for me, I kind of think of the ability to think and then speak quickly a little bit fluency, I guess. So when she answered that pretty deeply complicated question, so quickly, I marked her up on fluency.* (Production speed, fluency, turn-taking).

As noted by this rater, comments on speed were often made based on how quickly a test taker answered the interlocutor's question. Answering quickly was often a sign of comprehending the test question, and a quick response was furthermore a sign of stronger proficiency. In other cases, the opposite was true, with slower speaking being related to the test taker thinking about language and preparing a response, thus relating to lower fluency. Speed was not always related to language ability, however. In rater 20's view of the same sample 9, it related to affect:

> *I feel like part of why she's talking so quickly, because she's being anxious, and like just watching her eye movement and stuff, and kind of she reads more anxiously, but at the time, I was just like, wow, she's really going.* (Production speed, anxiety, gaze)

In this example, speed was seen as possibly excessive or unnecessary, and for this reason it related to anxiety. It was not the sole piece of evidence, though, as the rater also noted that eye movements factored into this decision.

The next most common paralinguistic feature in the dataset also related to language processing. Filled pauses, particularly "ums," "uhs," and so forth, appeared in 19% of the comments on paralinguistics by 14 raters. Filled pauses are a hallmark of developing fluency as they are assumed to reflect breakdowns

(Kormos, 2006; Segalowitz, 2010; Suzuki & Kormos, 2020), and raters noted it as such. For example, when rater 4 was describing sample 13, they observed: “I think with him saying, um, it just, it didn't give me, I wasn’t gonna give him like full fluency with him saying ‘um’ and sometimes stopping his speech” (Filled pauses, fluency). Raters also frequently noticed the absence of filled pauses as more fluid and proficient speech. However, comments also frequently overlapped with anxiety (34%) and confidence (23%):

> *[W]hen they were talking, you know, if they were stuttering a lot so, like people that had a lot of "uhs," you could tell she wasn't super confident and at ease. But once again, that could be much different things.* (Filled pauses, fluency, confidence, anxiety)

Here, rater 7 observed that the filled pauses produced by sample 12 led to impressions of a lack of assuredness, though noting that the behavior may be linked to other underlying causes. When speaking about sample 9, the same rater mentioned that these filled pauses “could have just been her trying to think of her answer.” In other words, this rater viewed these pauses as serving a cognitive role by creating space to think or a social-affective role conveying a lack of comfort in the testing situation. The use of filled pauses as continuers is also attested in the literature (Suzuki et al., 2021), as these alone do not predict fluency. Raters frequently mentioned that these pauses were, however, distracting, and may have caused problems with comprehension.

The third largest set of comments on paralinguistics was laughing, which frequently coincided with smiling. Laughing, appearing in 15% of the comments about paralinguistics, was mentioned by 15 raters and was generally seen as positive (except as one personal anecdote when nervous laughing was highlighted). Laughing coincided with comments about being at ease (21%), warm (15%), and expressive (14%). Rater 4 noted the appearance of positivity and warmth in sample 18: “And then I was saying that she was positive and warm, because she was like, laughing and smiling with the questions. So like, it felt more natural when she was talking” (Laughing, smiling, positive attitude, warm). It is notable in this example that the ability to laugh made the conversation seem more authentic to the rater. Rater 7 made a similar observation about sample 18 as well, noting that confidence and anxiety were impacted:

> *Confidence could be, you know, she's a little bit anxious about answering these questions. But at*

the same time, I gave her a five because laughing. You know, she, like, she was laughing she was… at ease… because that just showed, like, you know, I mean, people laugh when they're nervous, too. But I mean, it wasn't. It didn't seem like a nervous laugh. To me. It seemed like relaxing, getting into it. (Laughing, confidence, anxiety)

Although anxiety was anticipated because of the testing context, the rater noted that laughing helped make the candidate appear more relaxed, at ease, and eager to communicate her message. It also may have helped the test taker overcome problems with comprehension and to save face (Matsumoto, 2018; Pitzl, 2010). Overall, laughing helped to establish an evaluation of being involved in the conversation in a genuine and involved way.

Another notable category of comments regarding paralinguistics concerned tone and prosody, making up 13% of the comments across 8 of the 20 raters. Comments referred to vocal inflection, emotionality, tone, shakiness in the voice, being monotone, and stiffness. Comments regarding a wider range of prosodic features coincided largely with expressiveness (21%), warmth (13%), and happiness (13%). Rater 15 illustrated this observation about sample 13:

And like, he was saying it's sort of, you know, his tone changed a little bit because he's like, "Oh, well, how else will you decorate your home?" … I think I did score happier on, or higher on like, happier and it makes him seem, like more personable. (Tone-prosody, content, turn-taking, happiness)

In this particular case, the rater had previously discussed how the test taker's response was confusing, but it was the shift in tone that caused a more positive impression in terms of how approachable the individual was. The opposite also appeared to be true, as noted by rater 18 on sample 25 that "the reason why I put cold or inexpressive is I think that she's not really at the level of being able to convey that much emotion through speaking in English" (Tone-prosody, warmth, expressiveness, overall ability). Within this comment, the rater noted that the lack of prosodic features led to the test taker as being perceived as less approachable, while also noting that conveying warmth or emotion in the voice might have been related to overall language ability.

Tone and prosody, on the other hand, also aligned with anxiety in 13% of the intersections. These comments, although few, were due to comments about shakiness or timid-sounding voices. Rater 16 noted this about sample 12:

> *[I]t does sound like she has a shakiness in her voice. I can usually just recognize that. So I did put her higher on the anxious scale. But just because she's anxious doesn't mean that she's not like, she's not … competent.* (Tone-prosody, anxiety, competence)

The rater here previously noted shakiness as a feature of their own voice when nervous or anxious and thus related to the test takers in the testing context. However, as can be seen, the rater noted that this does not necessarily imply anything about the individual's language level.

Overall, paralinguistic features were somewhat less uniform in how they were perceived than gaze or mouth behavior, but these also varied quite substantially in form. In this dataset, while laughing and tone-prosody related to expressiveness and warmth, filled pauses most often related to anxiety and confidence. The remaining three paralinguistic features raters mentioned—audible breathing, backchanneling (e.g., mhmm), and voice volume—were mentioned far fewer times and will not be discussed here.

General face behavior

Comments concerning a general "look" on the face, the overall expressiveness of a face, or blank looks were coded as general face behavior. These broad comments comprised 11% of the comments on nonverbal behavior and were made by 17 raters (M = 2.65 comments per rater, SD = 2.03). These comments generally intersected with observations about anxiety (19%), confidence (13%), or expressiveness (13%). Rater 18, for example, noted that despite filled pauses, sample 17 was not perceived as anxious because of the combination of general facial expressiveness and use of prosodic features to convey emotion:

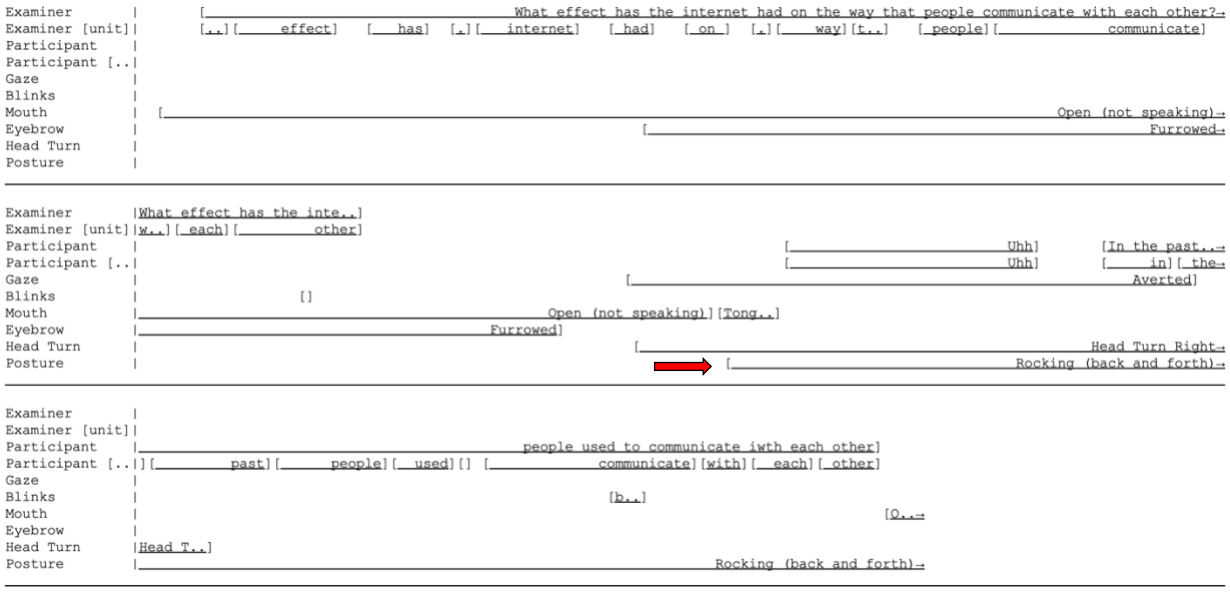
> *I think she's got more of like, a little bit more emotion in her voice and kind of like … her facial expressions. That's kind of like even though like she is kind of like stuttering and stopping and saying, "um", a lot. I think that's like less because she's nervous. So that's why I kind of had more at the at ease side.* (Face [general], tone-prosody, fluency, filled pauses, anxiety)

The rater in this example was keen to use a variety of features to better understand the test taker's possible underlying affective state. This was also true for rater 4 on sample 12, who commented that "for the hand movements I was said she was very expressive. So yeah, I would have given her a seven but her facial expressions didn't always match her body movements though" (Face [general], gesture, expressiveness). In this case, a lack of expressiveness in the face actually helped balance the rater's understanding of the test taker's affect from gestures. Expressiveness was often judged mainly through movement in the face, as noted by rater 6 on sample 16: "I rated him a little bit more inexpressive just because his facial expression doesn't really change much" (Face [general], expressiveness). Likewise, a lack of movement in the face often led directly to negative evaluations. When speaking about sample 19, rater 10 said, "So right away just facial expression, she already seems like, like very anxious and borderline to the point of being like uncomfortable and unhappy, stressed" (Face [general], anxiety, happiness). A more serious, stoic facial configuration has been reported as linking with anxiety in other studies as well (Gregersen, 2005; Lindberg et al., 2021). General face movements often served as evidence of question comprehension.

Posture

Features about the position of the body relative to the camera, as well as comments about the movement of the torso, made up 11% of the comments and were made by 17 raters (M = 2.60 comments per rater, SD = 2.11). The most common postural feature commented on was visible rocking or shaking (24% of comments by 10 raters), largely due to comments about sample 16. This behavior almost exclusively intersected with comments about anxiety (65%). In this test sample, after hearing the test question, the test taker almost immediately began rocking back and forth at the six second mark, which raters found visually salient. The behavior stood out partly because the behavior began so early in the interview, continued for 10 seconds, and recurred in the sample. The multimodal transcript in Figure 7.5 shows the onset of this behavior in context.

Figure 7.5
Sample 16 Rocking and Eyebrow Behavior



The raters unanimously found this behavior to be a sign of anxiety, as with rater 3:

> *I rated it higher for anxious because of the body rocking. Which, at least to me is like, it's very common body language is something that I'll see a lot. A lot of people will do as him when they're anxious*. (Rocking-shaking, anxiety)

Others were more willing to balance behavior across the interview, with rater 8 noting that:

> *[J]ust from the first impression that since he was going, like, back and forth, I thought he was anxious, not very at ease. … But I think later, he was anxious at first, but then later, he kind of like, calms down. That's why I initially rated it as at ease.* (Rocking-shaking, anxiety)

Having seen sample 16, some raters also commented on the lack of rocking or shaking when building a rationale for the relative calm or ease of a test taker. This was one of the few behaviors in the dataset with such a clear relationship to a judgement of affect. Whether this rocking was a sign of stimming, a rocking behavior typical of neurological conditions such as autism (American Psychiatric Association, 2013), is unknown. The relationship between behavior and rater training for neurodiverse individuals will be explored further in Chapter 8.

A second set of comments concerned leaning forward, made by 7 raters, and comprising 20% of

the comments on posture. Observations of leaning forward intersected most with engagement (29%) and attentiveness (24%). For example, rater 14 noted that further into the interview, sample 16 used his body posture to convey that he was engaging with the speaking test by leaning in: "And I noticed he was very attentive. When the guy would ask a question, he leaned right in and listened" (Leaning forward, attentiveness). A similar comment was made on sample 9 by rater 19: "And then how close she is and how quickly she responds, as well as the, in the middle of the question. She did one of those like she said, 'yeah', so I said she was relatively engaged" (Leaning forward, turn-taking, speed, active listening, engagement). As seen here, these comments state explicitly or implicitly that leaning forward was related with the test taker listening carefully to the examiner's question. This behavior generally aligned with positive impressions regardless of whether the leaning happened during moments of clear comprehension or comprehension difficulties, possibly because an attempt to understand a test question (rather than avoidant behaviors or answering without understanding fully) was always seen as a desirable behavior. Forward leaning behavior has been attested as a sign of involvement and rapport (Burgoon et al., 1984; Mehrabian & Williams, 1969), as well as engagement and listening comprehension (Jenkins and Parra, 2003; Neu, 1990). The topic of listening behavior, however, will be explored more in the next section.

Moving around, a very general postural behavior that made up 19% of the comments and by 5 raters, did not directly correlate with any given behavior due to its general lack of description. As seen in the comments on face behavior, some of these comments about general movement indicated expressiveness. Others, such as movement associated with leaning forward, indicated engagement and interaction with the test. Rater 5 noted that the test taker in sample 13 "felt engaged in the conversation like moving around. Like I was saying before, the head movements and stuff like that, and like he's interested in the question, actually, like, thoughtful about it" (Moving around, engagement, head). This type of synchronous movement with responses with the test was viewed positively and has also been noted as a positive aspect of performance in the literature (Jenkins & Parra, 2003; Neu, 1990). Other postural movements were seen as indicating anxiety, such as rater 5's observation about sample 12:

[S]he seemed kind of anxious just a little bit based on like, all the movement. It seemed like she was

kind of getting her thoughts out as they came into her head, which was probably like, why it seems so rushed. (Moving around, anxiety, thinking, speed)

Although this comment goes into little detail, the movement this rater observes may very well be an unsynchronized movement, which has often been linked to struggle in previous studies (Ducasse & Brown, 2009; Gan & Davison, 2011; Gullberg, 1998; May, 2009, 2011; Neu, 1990; Sato & McNamara, 2019; Thompson, 2016). A similar behavior, that of adjusting the body's position or shifting in the chair, made up very few comments but was often also tied to anxiety.

Finally, a straight or rigid posture, comprising 15% of the comments by 7 raters, and leaning back/slouching (10% of comments by 5 raters), made up the final set of comments. Like many of the behaviors, comments on straight posture intersected mainly with anxiety (21%), expressiveness (21%), and confidence (16%). This category of behavior included cases where the person was seen as positive due to their posture, such as rater 7's comment about sample 15: "Confidence would be how quick they answer each question, personally, and just like their body language. And like, you know, if they're like sitting up or if they're like, kind of hunched" (Straight/rigid, confidence, body language). The rater here again combined postural style with the speed or response and overall body language to formulate an impression of confidence. Sitting up straight was seen as confident, while being hunched over was seen as lacking confidence, but only if other behaviors suggested corroborated with the same affective interpretation. In other cases, leaning back or slouching was seen as being engaged and at ease, such as in sample 13, who visibly leaned back in his seat for a large part of the interview. Rater 9, for example, said of this test taker that "I think I noticed very much just he looked comfortable in the way that he was sitting not like tense and anxious" (Leaning back/slouching, anxiety). In some cases, sitting up straight contrasted with restless shifting around the seat or other body movements. Rater 5 noted that this lack of movement appeared to show a lack of expressiveness in sample 16: [H]e didn't really show too much emotion. … He didn't really use any hand signals when he talked …. didn't really I guess move around a lot. He was just there, maybe looked around a little bit, but he didn't like show any other type of expressive like behavior" (Straight/rigid, gesture, gaze, expressiveness). Again, this rater made their observation based on multiple body movements

when dealing with posture. Sitting up *too* straight, or too rigid, however, could also be a sign of anxiety, as rater 19 observed saying that in sample 29, "she appears like a very upright demeanor. I think sure her hands might be folded as well. Her shoulders are very, like close together. So she appears to be quite anxious to a certain degree" (Straight/rigid, lack of gesture, shoulders, anxiety). From this, straight or rigid posture and leaning backward could mean many things to raters depending on the overall observation of behavior in the individual. This rater mentioned the final category of shoulder movements, which was infrequent in the dataset but always related to anxiety.

Gesture

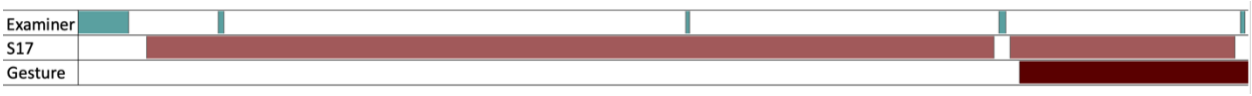
Comments about gestures—broadly speaking, any movement involving the hands—made up only 7% of the total comments on nonverbal behavior, mentioned by 13 raters (M = 1.75 comments per rater, SD = 1.59). Although relatively small, this figure is notable because the hands are generally not visible in Zoom recordings, including in this dataset. Nonetheless, seven of the samples did demonstrate behaviors that raters found salient. The most frequently mentioned gesture was the self-adaptor, which made up 43% of the gesture comments by 9 raters. The self-adaptors observed included hair touching, face touching, head scratching, and playing with the lips. These almost always coincided with judgements of anxiety (55%). Rater 15, for example, described two self-adaptors when rewatching sample 17:

> *She seems to like get more nervous towards the end because she starts, like, picking at her forehead and stuff. And at the beginning, she seemed like, just more comfortable. But I did notice that just because she started like fiddling like with her hair and stuff.* (Self-adaptors, anxiety)

Figure 7.6 is an annotation density graph of three lines of behavior in this sample. The top line represents the examiner's speech in teal, showing the test question followed by backchannels and one short follow-up question ("Why?"). The second line in light red is the test taker's speech, which continued throughout most of the sample here. The bottom line in dark red represents the appearance of self-adaptors. The rater's observation was consistent with the nonverbal annotations, as the test taker transitioned between a calmer demeanor to one with visible hand movement to the face. The picking and hair touching signaled to the rater that the test taker was experiencing anxiety as this part of the test unfolded. Indeed, self-adaptors are

often associated in the literature with coping mechanisms during times of stress in adults (Ekman & Friesen, 1974; Kikuchi & Noriuchi, 2019; Gregersen, 2005). In the context of an oral proficiency interview, this could possibly align with the general course of tests becoming more difficult as they proceed, possibly representing a spike in cognitive load.

Figure 7.6
Annotation Density Graph of Sample 17's Self-Adaptors



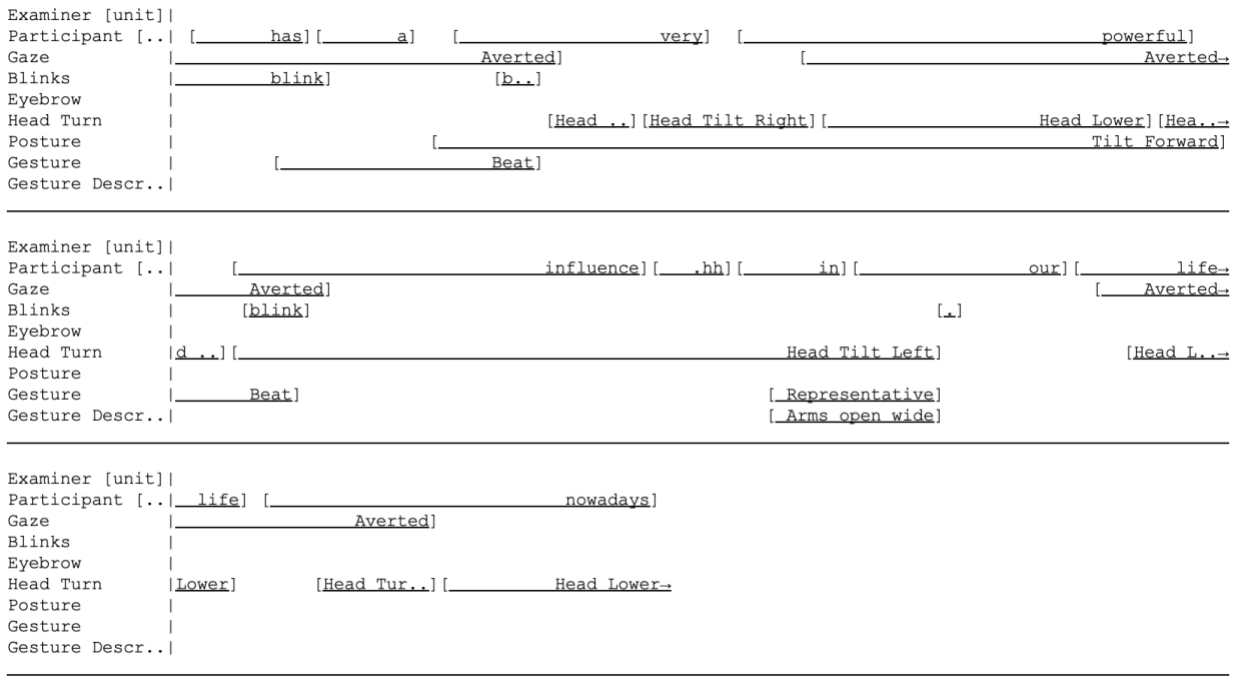
Representational gestures—iconic and metaphorical gestures that refer to some sort of object, event, action, or idea—comprised the next-largest group of comments on gesture (26%) being mentioned by 5 raters. These probably included beat gestures as well (as in the example below), but raters did not make a distinction between these. Fine grained distinctions between iconic and metaphoric gestures were also not possible in such a small sample, but raters made a clear distinction between desirable hand movements that aligned with speech (e.g., talking with hands) and those that were seen as more random and far less desirable (e.g., fidgeting). Representational gestures aligned with positive judgements of expressiveness (31%) and happiness (19%), and engagement (13%). For example, when describing sample 12, rater 3 mentioned:

> *[S]he was using her hands to talk … as someone who speaks with their hands, it is a little bit more expressive to me to speak with your hands because that's … an indicator that you're engaged in the conversation like expressing what you're thinking*. (Representational gestures, expressiveness, engagement, thinking)

This rater observed that using gestures during speech was something desirable, possibly helping to convey additional meaning through these hand movements. The sequence is presented in Figure 7.7, slowed down so that the gesture is fully visible. The test taker here was discussing the impact of the internet on online shopping. When she said, "has a very" and "influence," she used beat gestures to emphasize these semantic units. When she said, "in our (life)", she opened her arms wide, just slightly visible at the bottom of the video. This opening arm metaphorical gesture may be a sign of inclusivity, strengthening the meaning of

"our" as non-exclusive "we" (English only has one form of "we" that can include or exclude the second person interlocutor, only interpretable from context; Chinese uses 我们 for inclusive/exclusive "we" and 咱们 to indicate only inclusive "we"). Importantly, the rater noted that the use of these gestures (both beat and representational) is a sign of thinking about concepts (rather than language), an important dimension of McNeill's (1992, 2005) growth point hypothesis. Beat gestures may also be an important sign of prosodic control, revealing key information about the speaker's fluency and automatization of L2 use (McCafferty, 2006). The rater's observations led to viewing the test taker being more expressive and engaged. The co-alignment with speech and connection to topic content also appeared to make the test taker appear more proficient to the rater, aligning with Gan and Davison's (2011) findings.

Figure 7.7
Sample 12's Representational and Beat Gestures



As noted earlier, however, gestures need to align with other nonverbal behaviors to form a positive holistic impression of the test taker. Random movements, in contrast with representational gestures, were unanimously seen as negative and relating to anxiety. These were different from self-adaptors in that they had no noticeable form; self-adaptors were identifiable by their action (e.g., touching hair), while random

movements did not add a dimension of meaning to speech and were often unclassifiable. There were only 4 observations in the dataset. For example, rater 2 described sample 9's fidgeting gestures:

> *I remember thinking that she was pretty, she must have been pretty nervous at the beginning. She sort of relaxes later on, if I'm remembering this video correctly, but she's saying "um" a lot, she seems to be like a little fidgety.* (Random movement, anxiety, fluency)

In this example, it is likely the combination of filled pauses and random movement that lead the rater to form an impression of anxiety in this test taker.

The final category of gestures, that of lacking hand movement, made up 20% of the comments about gesture. The comments were generally split in this category. There were comments that noted the absence of fidgeting, which was seen as a positive attribute. An example of this type of comment was rater 6's observation of sample 13: *[He] wasn't looking around and adjusting and fidgeting a ton. So his voice seemed to remain like, I guess not shaky. So that's why I said, he seems to be pretty confident and not anxious* (Lack of gesture, shifting gaze, tone-prosody, confidence, anxiety). The rater here used multiple lines of evidence of the lack of behaviors that have so far associated with anxiety to form a positive impression of the test taker. This contrasted with the lack of desirable representational gestures, which was generally seen as negative, as noted by rater 5 about sample 16:

> *[H]e didn't really show too much emotion. So it was like everything else was just kind of in the middle. He didn't really use any like hand signals when he talked, didn't really I guess move around a lot. He was just like there, maybe looked around a little bit, but he didn't like show any other type of expressive like behavior.* (Lack of gesture, rigid/straight posture, expressiveness)

The rater explicitly noted that the lack of overall movement in both the hands and the body made the test taker appear far less emotive and less expressive. This did not lead to a negative evaluation, however, as the rater later noted that the test taker was likely focused on the task at hand rather than trying to communicate or express a certain idea. It is important to note that lacking co-speech representational gestures can be associated with lower proficiency (Gan & Davison, 2011; Gregersen et al., 2009)

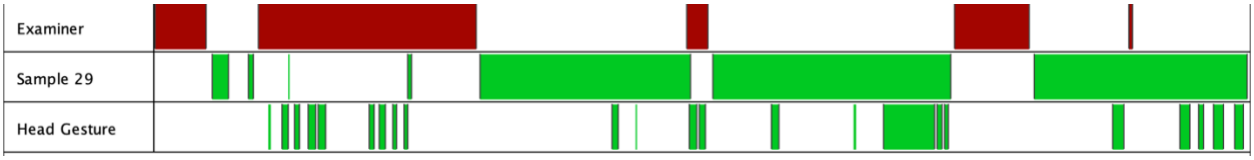
Head

Head movements made up a small category of two behaviors (6% of the dataset) observed by 13 raters (M = 1.50 comments per rater, SD = 1.76). The vast majority of these comments concerned head nods (74%). Head nods were positively associated with most judgements of affect, though notably attentiveness (18%) and engagement (18%). For example, rater 6 commented that sample 29's nodding was positive:

> *I think it was the nodding, and like, I guess just saying, "okay", as the interviewer was explaining, which made me rate both attentive and expressive, even though she did not understand, I guess, the question the beginning. That was her actively trying to understand it as it was being explained, so I think the attention was paid to the interviewer.* (Nodding, turn-taking, attentiveness, expressiveness, comprehension, active listening)

The rater observed in this extract that active listening, both in the form of head nodding and verbal backchanneling ("okay") were important tools to show active visible and audible engagement in the communicative event. Figure 7.8 illustrates the extensiveness of this test taker's use of head gestures, most of which were nods, though these included one extended nod and two head shakes. The rater saw this behavior as positive despite the ongoing comprehension difficulties the test taker had with one unfamiliar word at the beginning. Sometimes nodding added to the level of confidence someone exuded, such as rater 7's comment on sample 16: "Also, just something small, head nodding once he finished his sentence. Like, 'I know what I said. I'm confident my answer. I didn't think I messed up and on anything.'" The test taker's nod in this context was seen as both an affirmation and a skillful turn-taking device to give the floor back to the examiner. This interactional skill was almost always seen positively.

Figure 7.8
Sample 29's Head Behavior



The second category of head behaviors concerned head turns, which only comprised 8 comments (26%, by 8 raters). These comments were largely heterogenous, ranging from "moving the head back and

forth" (rater 8, sample 13), "turning her head" (rater 5, sample 17), "bobble head movement" (rater 10, sample 13), "constant little head movements" (rater 10, sample 29), "head bob" (rater 20, sample 16). In each of these examples, these movements were seen as positive, indicating some level of comfort, ease, and expressiveness in the display of movement. There was only one case where "mini head movements" (rater 5, sample 9) combined with shifting gaze led to an impression of anxiety. In all cases the movements were factored together with other behaviors to form a holistic impression of the test taker.

Eyebrows

Of all the specific areas of the body mentioned by raters, the eyebrows were the least frequent. There were only 11 comments in total made by only 6 raters, representing only 2% of this dataset ($M = 0.55$ comments per rater, $SD = 0.94$). For this reason, the discussion will be brief. The most common comment about eyebrows was their general movement, making up 6 of the 11 comments. Some raters, such as rater 6, found eyebrow movements to represent engagement and interactiveness, such as with sample 12: "And like eyebrows in general, and eyes are moving up and down rather than just straight face. So I said that was engaged and interactive" (Eyebrow movement, engagement, interactiveness). Others, such as rater 12, found them to be a visual indicator of thinking, as in sample 17:

> *[S]he was expressive with the eyebrows, but not so much on the face, which is why I thought colder … But I do remember like, she is thinking very hard, maybe not as expressive, but the eyebrows are like, the eyes are like okay, she's definitely thinking, she's definitely having, you know, like, visual reactions.* (Eyebrow movements, expressiveness, gaze, thinking)

In this example, the movement of the test taker's eyebrows was seen as a positive indicator of internal cognitive processes, but not of overall positive affect because they did not align with over visible behaviors. These findings appear to align with the positive impact of eyebrow movements on fluency (Kim et al., 2023; Tsunemoto et al., 2022), as these movements were able to mark prosodic stress in ways that aligned with fluent speech.

Comments on furrowed and raised brows made up the remaining 11 observations. Furrowed brows (3 comments) were associated with comprehension problems. For example, rater 15 observed the following

about sample 16:

> *[A]lready he seems confused, like, right when he started talking. He like furrowed his brow. So I remember thinking that even before he started talking, I was like, it seems like he's gonna have trouble from just his reaction to being asked something*. (Furrowed brow, comprehension)

The behavior rater 15 mentioned can be seen in Figure 7.5. The test taker began furrowing his brow just after hearing "What effect has the internet had", before the completion of the examiner's question. Rater 15 used this information to preemptively prepare to *hear* a repair sequence, even though there was no such repair sequence in this sample. Raised brows, of which there were only two comments, were associated with attention, such as the following comment by rater 7 about sample 9:

> *They're engaged, at ease. … Body language, you know, she's facing the camera, obviously, her back is straight. She's not frowning or anything. Okay, her eyebrows are pretty raised. She just looks ready, attentive*. (Raised brows, engagement, body language, posture, furrowed brows, attentiveness)

In this extract, the rater mentioned several lines of evidence that the test taker was relatively at ease and engaged. The rater indicated that the raised brow aligned with these behaviors, but seemed to imply that the eyebrow raising might be a sign of anxiety. It could be, then, that raised eyebrows were an interactional device to signal to the examiner that full comprehension was not reached, as was the case with furrowed brows.

General body language

The final category was the most general of all and included observations about any movement of the body without specific details. There were only 10 of these comments by 6 raters, making up 2% of the dataset (M = 0.50 comments per rater, SD = 0.83). Raters generally referred to this category using the specific term "body language," but also included "change of mannerisms," "physical movements," "how they were in the video," and "their looks." These comments were heterogeneous in nature providing evidence for a range of affective states, such as anxiety, confidence, engagement, and others. Two comments also indicated that body language provided evidence of an individual's overall ability *before* the

test taker started speaking. Rater 11, when asked why they initially thought that sample 13 "was not going to be good," said:

> *I don't know, I feel like I would have had to base it off looks, which is something I didn't do throughout the rest, the whole thing. But I felt like I have to mention it … as soon as I was like, "Oh, he's not going to be good," I just remember feeling like that, and literally all I did was see his face. So it had to be maybe about the way that he was looking about it first.* (Body language, face behavior, overall ability)

In this example, the rater later recalled that the test taker performed well on the test, and by the end the overall judgement was positive. It may be the case in other situations, though, that observing nonverbal behavior prior to hearing speech colors the raters' interpretation of language proficiency throughout a particular test.

Behavior, affect, and language proficiency

In this study, I had anticipated patterns emerging about nonverbal behavior and proficiency judgements. The reality, however, was more complex because judgements of affect and nonverbal behavior were often intertwined. In fact, while there were some behaviors that were generally associated with positive judgements of affect (e.g., mutual gaze, smiling, use of prosodic features) and others with negative judgements (e.g., shifting gaze, filled pauses, self-adaptors), broadly speaking, raters did not often indicate explicit ties between specific behaviors and evaluations. Instead, they considered nonverbal behaviors as a context-bound cluster of phenomena that interacted with linguistic features and the content of utterances to create a sense of affect, which occasionally impacted impressions of language ability. Raters considered multiple lines of evidence when forming their impressions and balanced these to arrive at their judgements.

While specific behaviors were not found to impact language proficiency, comments were made about the utility of nonverbal behavior when forming impressions of language. For example, rater 1 made the observation that:

> *I think it is important to kind of integrate facial reactions and like, not reactions, but expressions on how they're reacting physically to statements as a reflection of fluency, or maybe even*

vocabulary, like whether or not they're understanding what you're saying, is a reflection of their own lexical knowledge.

This statement indicates that facial behavior has the capacity to provide extra information about an individual's linguistic competence and language processing that would otherwise be unavailable to someone just listening to the recording without visuals. Rater 12 made a similar observation about the use of behavioral information when judging language:

Just a lot more movement, a lot more thinking. I think with her when she said, it was her trying to think. The eyes too, the eyebrows, definitely more expressive. I think she smiled a couple times. But I think maybe it was like, the accent, so I couldn't definitely tell how to rate her, a weaker vocabulary, or weaker grammar. I think it was just easier with her specifically to see her face.

From this extract, the rater did not associate the test taker's positive nonverbal expressiveness with any concomitant positive features of language. In fact, the rater noted that the language skills, displayed through pronunciation, vocabulary, and grammar, appeared somewhat less proficient. However, the presence of nonverbal behavior aided in their final decision, thus providing additional information that would have been unavailable in an audio-only format.

In one extract mentioned previously, rater 11 noted that the presence of nonverbal behavior might also create an impression of language ability even prior to hearing the individual. Although the rater noted that the look on the test taker's face conveyed a negative impression in terms of language ability, they also said that this was uncommon. Nevertheless, it was the first instance of behavior that elicited this type of response from the rater and shows how certain visible phenomena may impact proficiency ratings.

Comments specifically relating nonverbal behavior to proficiency were, however, few. To discover deeper meaning in the dataset, I used axial coding of the already coded utterances to uncover other patterns. After repeated readings of the dataset and analyzing intersections of behavior and language, I devised, revised, combined, and distilled sets of thematic codes that represented patterns within the dataset related to language proficiency judgements. The final set of thematic codes and code counts is presented in Table 7.4, and each code is discussed separately below. For the sake of coherence, the order of the discussion is

thematic rather than in order of pattern strength.

Table 7.4
Thematic Codes

Theme	Count	%	Ext
Multimodal assessment of listening comprehension	99	46	19
Assuredness impacts perception of proficiency	46	21	15
Approachability moderates perception of comprehensibility	42	19	15
Adaptability moderates impact of breakdowns	30	14	14

Multimodal assessment of listening comprehension

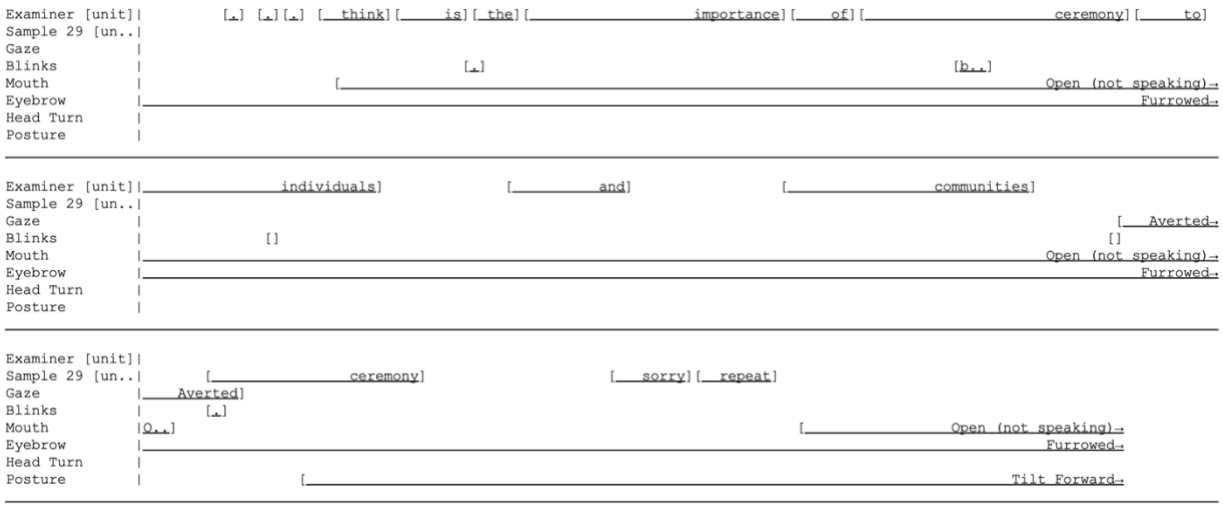
Although raters were not asked to rate or even consider whether the test takers understood the examiners' questions, comprehension factored all 20 raters' decision-making processes. On its own, it was the fourth most commented code, following speech content, fluency, and vocabulary. Comprehension of test questions emerged as an early signal of a test taker's overall proficiency, and raters frequently used this to assess language skills. However, unlike speech, listening comprehension is not directly observable. Raters used multiple lines of evidence, most importantly using nonverbal behavior, but also receipt tokens such as *oh, ok, yes* (Heritage, 1984, 1998) to assess listening comprehension. This multimodal assessment of comprehension emerged as the most common pattern in the dataset, occurring 99 times across 19 raters (M = 4.75 times per rater, SD = 3.70). For example, rater 10 described evidence of sample 29's breakdown:

> *So right away just facial expression, she already seems like, like very anxious and borderline to the point of being like uncomfortable and unhappy, stressed. But then I did notice that like, right away, she tried asking "ceremonies?" which a lot of people just kind of like, let it go until like I said before, like it was just like that uncomfortable that it's clear, they didn't understand.*

The rater in this example noted that the test taker's facial expression indicated quite negative affect, signaling discomfort. This discomfort was interpreted as an immediate indication of an underlying problem, even before the problem was made explicit. Figure 7.9 shows the behavior that led to this interpretation. The test taker held a furrowed brow from even before the test question began, and shortly after the question started held her mouth in an open position without speaking. This combination of behavior likely conveyed a sense of anxiety and unhappiness, though the mutual gaze held throughout indicated the test taker

remained focused and engaged. The explicit communication of non-understanding and identification of the trouble source appear 850ms after the test question ended. The test taker used a restricted repair initiation ("Ceremonies?;" Dingemanse, et al., 2016) to convey that she had not understood this specific lexical unit. The rater in this example appeared to imply that identifying the trouble source with this restricted repair initiation was favorable, but combined with the behavioral and affective evidence solidified their assessment of listening comprehension.

Figure 7.9
Sample 29's Comprehension Breakdown



Assessing listening comprehension frequently associated with judgements of overall competence. If competence is understood as being able to complete a "task" appropriately, this is a logical connection, as answering a question contingently would satisfy its requirements. Rater 14 made this observation about sample 18:

> *Here's when I was like, Oh, she's very fluent, because her words are flowing right out, and I could understand everything… when she started, she's smiling, and she's attentive and understood the prompt. So I also said she was competent, because she did the right thing.*

Competence also related to judgements of overall language ability. Rater 7 commented that "[h]onestly, for the competent/incompetence. I viewed that that one as like an overall, like an overall scale." These comments show the importance of assessing listening comprehension through multimodal evidence, as it

allowed them to form a holistic impression of the test taker's language proficiency.

Assessing listening comprehension also appeared to give raters specific information about the test taker's vocabulary knowledge. For example, rater 9 commented on this issue with sample 18:

> *This one I do remember thinking the way she hesitated, looked around. Definitely seemed like she didn't understand what the question was, which I think, I mean, I've been saying vocabulary a lot based off of how they've been responding. But I think that's also what they're able to understand. So her vocabulary score was, it wasn't low, but it wasn't super high.*

Instead of basing their decision on actual vocabulary produced in the test, the decision this rater arrived at involved vocabulary that was not understood and not produced. In other words, this rater and others used deficiencies as a line of evidence rather than basing decisions on what the individual was able to do. The rating scale used in this study likely lent itself, at least partially, to this type of behavior as decisions were mostly binary (positive/negative), and the raters were untrained on any understanding of language proficiency. Contemporary rating scales for L2 speaking tests, on the other hand, generally draw on can-do descriptors or descriptions of language use at target levels, and these do not generally include a description of language deficiencies (see e.g. Council of Europe, 2020). Whether trained raters engage in similar deficiency-based thinking is a question that remains and is beyond the scope of this paper.

Adaptability moderates impact of breakdowns

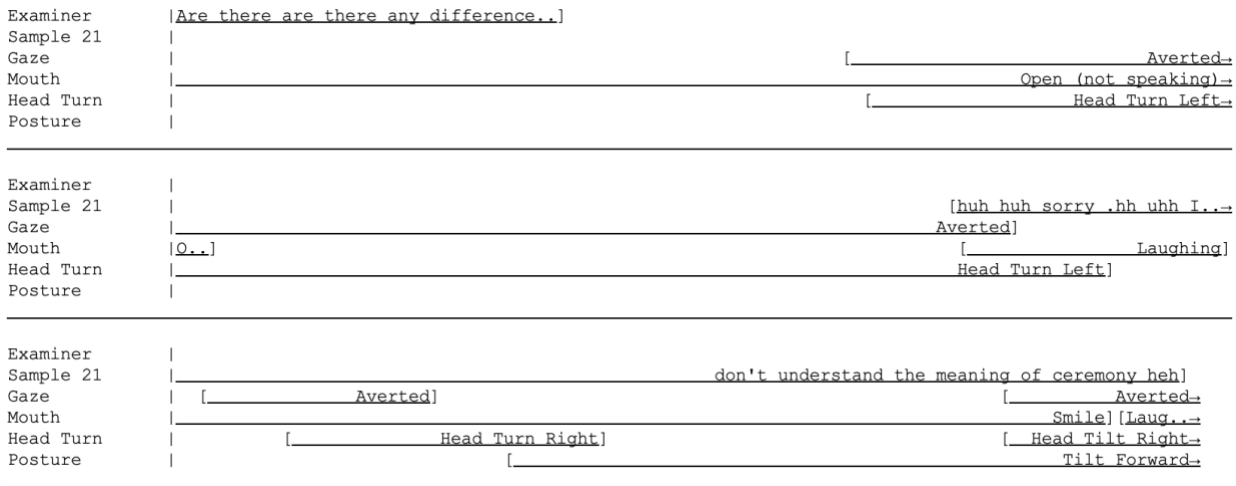
The raters in this study always viewed breakdowns in understanding as negative. As discussed in the previous section, it impacted overall impressions of language proficiency as well as impressions of lexical competence. This negative impression, however, was sometimes attenuated by the behaviors and affect the test taker exhibited during the breakdown sequence. These attenuating behaviors included verbal aspects of interactional competence, the test takers' turn-taking moves, repair initiations, relevance-contingence of their responses, and active listening. Many comments also focused on identifying whether the test taker was thinking about content or language, and these decisions were primarily made based on the nonverbal behavior test takers exhibited in their facial expressions. Likewise, affective stances such as a desire to communicate, engagement, and focused attention also factored into whether the impact of the

breakdown was stronger or weaker. Together, this cluster of interactional, cognitive, and affective behavior provided the raters an impression of their adaptability, or "how a candidate copes in a novel or challenging language situation in real time" (Harding, 2014, p. 192). Adaptability mediated evaluations of test takers in light of breakdowns, with more adaptable test takers benefitting. This pattern emerged 30 times across 14 raters (M = 1.50 times per rater, SD = 1.61).

Adaptability was sometimes characterized by how personable a test taker was during the breakdown sequence. For example, the test taker in sample 21 failed to initially understand the test question in Figure 7.10. She displayed this nonunderstanding 880 ms after the end of the question by averting her gaze and turning her head left. She held this behavior until she initiated an other-initiated repair by laughing, re-establishing mutual gaze, and returning to a neutral head posture. This display of positive affect, attention, and engagement may have established her comfort in asking for help, making her appear more personable. As she began verbally requesting repair, she smiled while leaning forward toward the camera to continue showing her engagement and attention to the examiner.

Figure 7.10

Sample 21's Breakdown and Repair Sequence



Rater 6 noted the positive impact of these behaviors because she remained warm and friendly, but also attentive, throughout her breakdown, despite the potential stress of the breakdown:

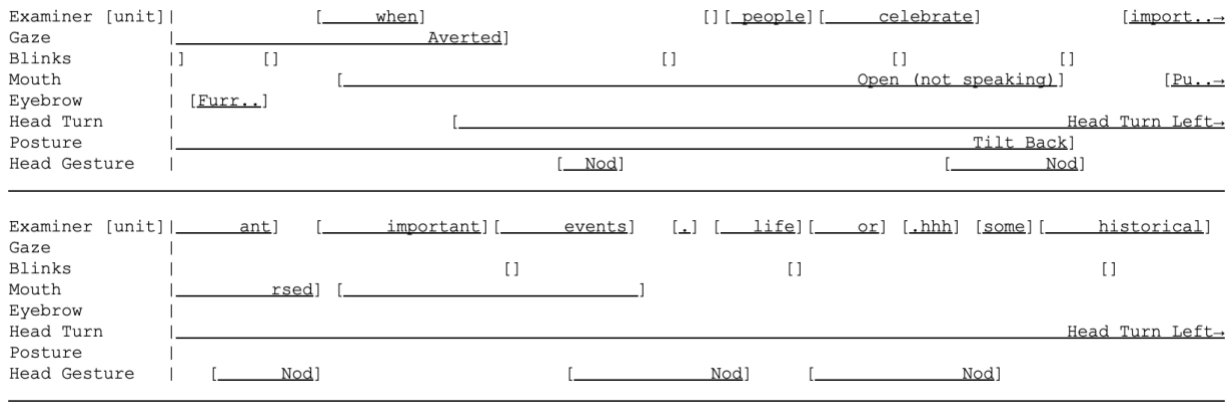
I remember saying warm because even though she was confused, she didn't seem overly anxious

> *about it. Like she laughed about it and asked the interviewer to repeat the question rather than like, eyes darting around, eyes was getting bigger. Plus she was attentive and expressive as well with a smile and overly seems happy or overall seemed happy despite the confusion.* (Laughing, warmth, comprehension, anxiety, repair, shifting gaze, eyes growing wide, attention, expressiveness, smiling, happiness).

In this scenario, despite the ongoing comprehension difficulties she faced, her positive attitude by smiling and laughing throughout the repair sequences resulted in an overall positive impression. Likewise, the rater noticed that the test taker kept mutual gaze with the examiner during the repair sequence without her gaze drifting excessively. The rater viewed these behaviors as oriented towards communication as she was paying attention and trying to collaborate with the examiner towards responding.

Likewise, behaviors that indicated active listening also factored into the perception of adaptability during breakdowns. For instance, the test taker in sample 29 initiated a repair sequence after hearing the unfamiliar word "ceremony," which was shown in Figure 7.9. After this, the examiner provided a clarification sequence, shown in Figure 7.11. As seen here, after the examiner began speaking, the candidate relaxed her brow and soon after her body posture, leaning forward from a backwards lean during her repair request. She then showed that she was actively listening not by smiling or laughing but by reestablishing gaze with the examiner and nodding just after key words the examiner repeated (when/celebrate/important/events/life). As illustrated in Figure 7.8, this test taker used nods frequently to backchannel and show active listening with the examiner. Her slight turn to the left may have also been a head gesture of directing her ear to the camera, which could have been a sign of active listening.

Figure 7.11
Sample 29's Display of Active Listening During Clarification



Rater 7 noted the head nodding and mutual gaze in this sequence, relating them both to active listening:

> *Just something little, her head nodding, she was attentive. She was under... She was trying to understand what he was trying to say. I mean, there's that you can see it, kind of see how, that goes on with facial expressions. But yeah, and a little bit of confidence to my opinion, because, you know, she's not in her head. And, you know, obviously she's looking at [the] computer.*

Breakdowns in communication were almost always seen as leading to a loss of confidence, but the rater here noted that the test taker's adaptability helped her regain some of that confidence. Again, being attentive was seen as vital in this moment of breakdown and integral to her adaptability.

Even in cases where comprehension resolution was not fully reached, some raters still reached a positive evaluation of the test taker due to their adaptability during their response. For example, sample 15 failed to comprehend a question about rewards. Instead of answering the question about rewards in general, she responded with a personal anecdote. Rater 8 detected the non-contingent response, but still arrived at a positive impression of the test taker.

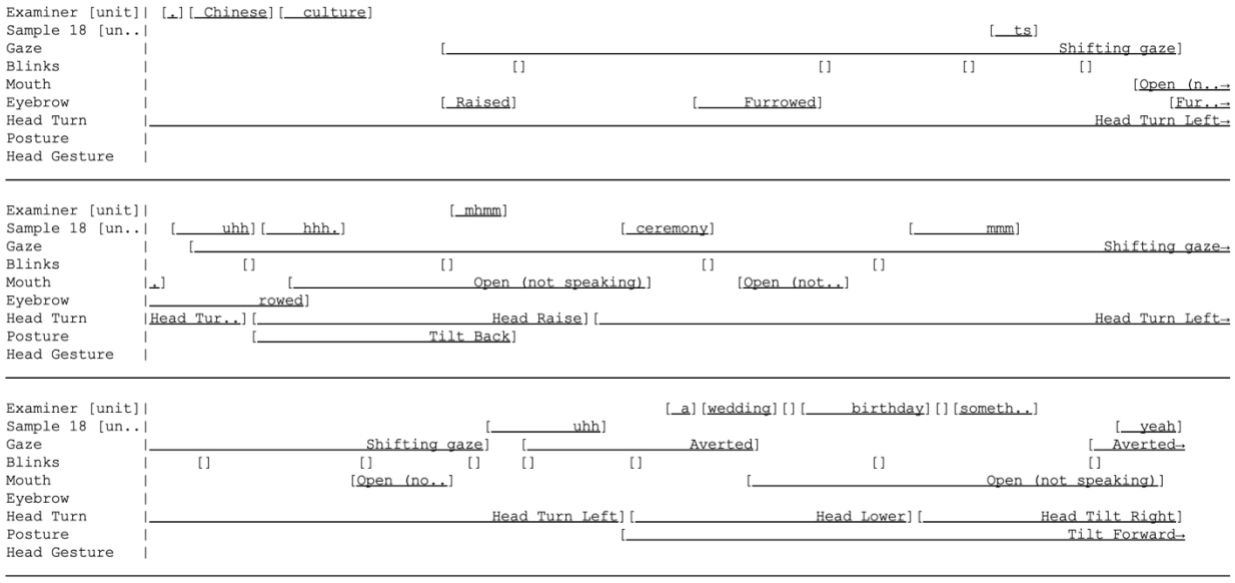
> *I didn't totally think that she* [comprehended the question] *at this time. Okay, just because she's talking about herself when the speakers asking about children and their parents' relationships. I still think she's at ease even though she doesn't really look like she knows what she's talking about*

yet. She's still calm. I think she's not like anxious, looking around, playing with her hair or anything like that. That also goes back to confidence as well, because she's still. Like, she looks confident in what she's saying. Even though what she's saying makes no sense to the question, but that's okay. We'll get there.

Despite providing a somewhat irrelevant answer to the question, the rater noted the general feeling of calmness of the test taker. She adapted to the communicative event despite not fully understanding the question, thus showing a desire to communicate. Her gaze was held on the examiner and she did not use self-adaptors during her response, reinforcing this impression of ease and confidence. Confidence for this rater was an important sign of overall language proficiency given the comment "but that's ok. We'll get there." This echoes a finding from Jenkins and Parra (2003) in which an off-topic response was perceived positively in a similar manner: "[Alejandro's] use of nonverbal features during his inaccurate response were perceived as confident behavior and had a positive effect on the evaluators" (p. 98).

On the other hand, breakdowns in comprehension combined with a lack of adaptability resulted in negative evaluations overall. For example, in sample 18, the word "ceremony" also led to a breakdown in understanding. The test taker's behavior, however, was far different from the earlier samples and showed much less adaptability. Immediately after the end of the test question, the test taker averted her gaze by looking side to side repeatedly, maintaining this shifting gaze until the examiner began a confirmation sequence. She initiated a repair, but indirectly by repeating the word "ceremony" with a falling tone, which was not a direct clarification question. The word was surrounded by filled pauses. She did not smile and showed little positive affect. Instead, she held her mouth slightly open, which may have conveyed confusion. Behavior showing active listening was also absent as there were no nods during the examiner's clarification sequence or after reaching some resolution at the end when she says "yeah." This sequence is displayed in Figure 7.12.

Figure 7.12
Rater 18's Breakdown Sequence



Rater 19 noted this relatively unadaptable behavior in the following extract. Here, the rater pointed out that the test taker's repair sequence was not explicit, which provided evidence about her communicative intent:

> *[S]he wasn't necessarily willing to communicate that she didn't understand ceremony if that is the direct issue that appears to be. I said that she wasn't quite confident. I noticed because previous people, like asked what ceremony meant or like, they didn't understand the word and they communicated that where she's kind of just trying to work through it and find the meaning. And then also quite expressive, because you can tell she has kind of a confused look. And she's her eyes are darting trying to find the answer.*

The rater noted that the candidate's reluctance to admit to the breakdown indicated a certain lack of confidence. Combined with an observation about the overall facial behavior, indicating confusion, and the shifting gaze, the examiner deduced that the test taker was thinking about language rather than content. This type of struggle likely provided more evidence of the lack of confidence, which overall was a negative evaluation.

Approachability moderates perception of comprehensibility

Comprehensibility was one of the scale categories for raters, and as such it was the fifth most commented feature with 228 comments, even more than grammar. I anticipated that comprehensibility would be easier for the raters to interpret than pronunciation yet rated similarly. The raters indeed scored this category on whether they understood what the test taker was saying, but because comprehensibility is broader than pronunciation, the raters also referenced elements of fluency, vocabulary, grammar, content (amount of talking, breadth of knowledge, relevance, truth, etc), together with pronunciation. Thus, this measure is not a clear representation of pronunciation alone. For example, when describing the ease of understanding sample 29, rater 2 mentioned:

> *It gets a little hard to understand. After she says, "they choose that day to get married." From that to this point is a little bit difficult to understand. Like it was hard for me to sort of figure out what she was trying to say… I think that it's a combination of like, misused grammar and accent, I do think the accent plays a factor here too. But I just think they're not. It's not a fluid conversation. Very, it's very choppy.*

Rater 2 mentioned a combination of grammatical errors, accent, fluency, and possibly some confusion about the content of the test taker's speech all as factors that made understanding her difficult. Accent was not always a main factor in decisions, however, as rater 13 noted when asked about whether their observations about accent impacted any of their scoring decisions:

> *I don't think it really does, because I think, like, as long as you can convey the idea, the idea is mostly conveyed through, in my opinion, grammar and vocabulary, and how well you're able to express yourself.*

Again, grammar and vocabulary formed the majority of the rater's understanding about comprehensibility, as the key evidence appeared to be idea generation. Content, then, was a main focus. Other raters echoed this, saying that they anticipated hearing an accent because they knew in advance that the speakers were second language learners, and for this reason accent played less of a role. These findings were in line with research on the differences between accent and comprehensibility in SLA (Munro & Derwing, 1995;

Trofimovich & Isaacs, 2012).

One pattern that emerged amongst these comments about comprehensibility was that test takers that were more difficult to understand sometimes benefitted from behaviors showing greater *approachability*. Approachability in this case arose largely from being seen as personable: having a positive attitude, being expressive, and actively engaging with the examiner. There was some crossover with adaptability in the sense of appearing at ease and confident, but resolving breakdown was not a key element in these comments. This pattern appeared 42 times across 15 raters (M = 2.10 times per rater, SD = 1.86).

For example, rater 1 explicitly mentioned affect and nonverbal behavior as key elements making the test taker in sample 8 more comprehensible:

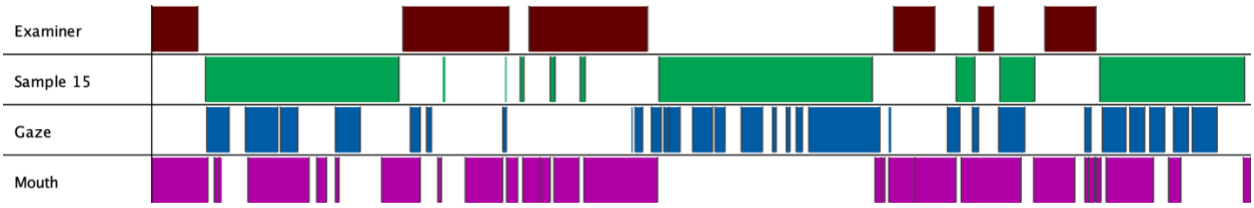
> *I was just thinking about it's more of an affect thing, kind of able to laugh off his mistake it's a nice little moment… the stuttering obviously made me think that maybe it's not... his fluency isn't quite there. He's not able to get to his mouth what's in his brain. But his presentation is really not bad. That's why this was interesting to me because his grammar and vocab and his overall fluency is really weak but his pronunciation is pretty comprehensible.*

Rater 1 noted that despite weak fluency, vocabulary, and grammar, the test taker's pronunciation was largely comprehensible. The rater appeared to attribute this, at least partially, to the test taker's affective stance during the test, taking a more relaxed approach and being willing to show a good-humored nature by laughing and recognizing his mistake.

Likewise, sample 15 was often discussed as an example where there were multiple problematic moments, including a breakdown in comprehension and speech that was somewhat irrelevant at points. Nonetheless, this speaker was very approachable during the entire interview. She smiled, showing a positive attitude, and she often maintained mutual gaze with the rater while he spoke and clarified her misunderstandings, showing engagement. A gaze density plot for the entire interview is provided in Figure 7.13, which shows averted gaze in the gaze tier, and mouth movements in the mouth tier. The test taker spent 25% of the overall sample time smiling at the examiner. It is also notable that while the examiner was

speaking, the speaker almost always maintained mutual gaze.

Figure 7.13
Sample 15's Smile Density



This approachable stance was noted by multiple raters. In the following example, rater 2 made the observation that her positive affect impacted how communicative and thus comprehensible the candidate appeared:

> *I think she was one of the first ones where it seemed like, you can meet her on the street and have a conversation with her… It's very it's natural seeming. She's not always perfect, but she's really like, her point comes across, which I think is more important than being perfect. She's very communicative… They're really expressive. But the vocabulary is not great. But they do get their point across. So like, people could understand them… So I do think that I rated her better, because she was able to get our point across and was very engaged and focused.*

Again, in this comment the rater noted weaker elements of language proficiency, but her expressive, engaged, and focused demeanor helped her to convey her message effectively. Maintaining this stance was then an important part of being perceived as comprehensible. Rater 10 echoed a very similar sentiment:

> *So I remember thinking about her personality, I don't know how much like to relate this to like the language skills, but her personality is very easygoing, comfortable. And then to that, like she's able to kind of express it. So she might have an easier time communicating with, like, even if someone else might know English better than her. I think she's very expressionate* [sic] *in the way how she moves her face and stuff. So I think she might have an easier time communicating.*

In this example, the rater explicitly noted that communication would be facilitated by her approachable stance, even despite language difficulties.

Lacking expressiveness, positivity, and engagement likewise factored into decisions that ultimately

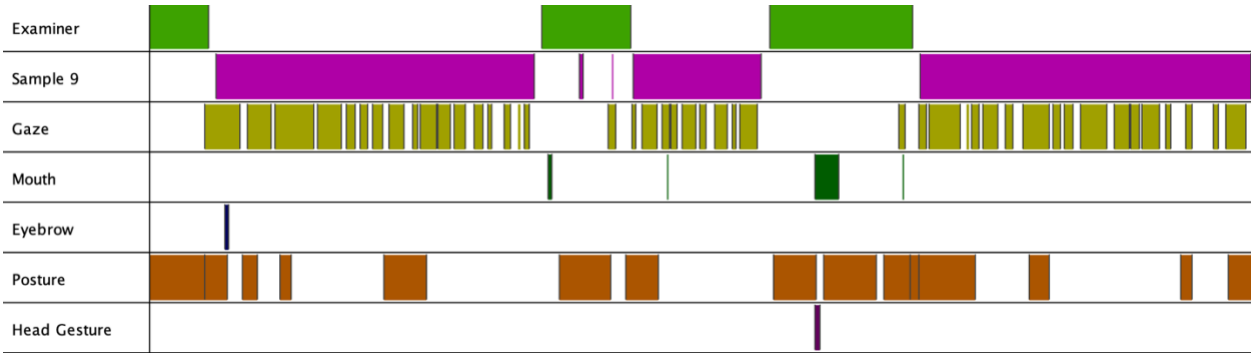
resulted in evaluations of lower comprehensibility. Namely, lacking expressiveness in the voice by speaking in a monotone manner was seen as less comprehensible. Rater 1 noted this about sample 9, saying that "that little chain right there, I remember feeling that I think this is where I really scored for comprehensibility, because, like it's all so monotone." As seen in the annotation density plot in Figure 7.14, this test taker's demeanor likewise lacked indications of warmth and engagement through mouth movements, eyebrow movements, or head gestures. Her mouth movements included holding her mouth open for three seconds in the middle of her response and three quick moments where she licked her lips. This test taker conveyed a sense of disengagement to some raters due to her gaze and facial behavior. This can also be seen in Figure 7.14, where there were frequent interruptions in her mutual gaze throughout. Rater 17 mentioned that this behavior made it difficult to follow what the test taker was saying:

> *She's like, looking everywhere. And it's not like she's holding it, and then maybe moving. It's like, bam bam bam. And like, it was like everywhere. Like, anxious or not quite sure where she's going with her sentence.*

In other words, her behavior was seen as rushed and unnatural, thus not connecting with the examiner, and creating an approachable stance. Rater 4 echoed this, saying her lack of engagement, observed through eye movements, a relatively static facial expression, and monotone delivery, made her discourse distracting:

> *I was thinking about how much she was saying "um" and looking around… I guess it was just distracting. And then it was also like a break every time she was saying a word, which was, like, distracting me from what she was trying to say… I felt like she wasn't very engaged in the question… Um, facial expression, um, and then like, no inflection in the voice, like she's just talking the same way the whole time.*

Figure 7.14
Annotation Density Plot for Sample 9



Assuredness impacts perception of proficiency

The final pattern of comments directly related to perceptions of language proficiency. Raters frequently mentioned that assuredness—greater confidence and lower anxiety—frequently aligned with their perceptions of overall language ability. This pattern was in reality the second most frequent, occurring 46 times and being made by 15 raters (M = 2.30 times per rater, SD = 2.39). In some cases, this perception as quite broad and holistic, such as rater 2's comment about sample 15 (who likewise appeared more personable and approachable in the previous section): "I think that her seeming comfortable, made me want to rate her as more fluent." The rater appeared to refer to the broader meaning of fluency here as language ability (Lennon, 1990). Simply appearing more at ease led the rater to a greater estimation of her language ability. Likewise, confidence led rater 4 to a broad, holistic impression of ability of the same test taker:

> *I think I instantly remember her sounding really, really good? I think I remember her face. I remember [her] sounding really good. I gave her a seven for being, me being able to understand her. And so she was calm and confident and warm. All good scores.*

This rater formulated a very strong impression of the test taker based on her confidence, relative lack of anxiety, and also her warmth. This observation also relates to her comprehensibility, which was noted in the previous section.

Some raters mentioned that nonverbal behavior led to an understanding of assuredness, and assuredness then led to an overall understanding of language ability. Rater 7 referred to ability as competence when describing sample 25:

Right there, I mean, that's where she kind of lacks some confidence in her answer. So I mean, that's why I didn't give her the full confidence. Her facial expressions, she's not super expressive. I mean, you can see she's kind of like looking off, you know, she's definitely trying to think of her answer. So I mean, that's where, you know, somewhat competence and incompetence comes into. Honestly, for the competence/incompetence, I viewed that that one as like an overall, like an overall scale. Like, is she competent at speaking in English? Or is she incompetent at speaking English? So that's how I viewed that slide on, you know, obviously, I gave her a kind of low score, because I mean, once again, I could understand what she was saying, but feel like each of her sentences that she was trying to speak, she was just like, jumping from word to word.

Here the rater started with an impression of confidence based on facial expressions and averted gaze. This lack of confidence from a lack of expressiveness and attention led the rater to understand that the test taker was struggling. The rater mentioned that she was thinking, and once she started speaking, fluency problems became evident through the lack of connected speech. However, the judgement of lower competence appeared to take place prior to any observations about language and was based purely on what the rater could see in the test taker's demeanor.

In other cases, assuredness was deduced directly from nonverbal and verbal behavior, but the relationship with overall proficiency was not necessarily causal. Rater 20 commented that for sample 17:

Yeah, there's also just, I mean, generally there's no sign of like, lack of confidence or anxiety. If she looks away or hesitates, its because she's thinking. I detected very little doubt of her own abilities. Or, there were very few like awkward pauses. Not much awkward, like mouth movements or, you know, eyes darting around for no reason, things like that. And she stays very centered. Very, she seems very confident when she expresses herself and her language skills are very high. And again, that kind of combination of confidence and language skills gave made me give her high competency.

Although the rater did not explicitly mention a direct relationship between the two, the characteristics of being confident, at ease, and stronger language were all related, leading to a strong overall impression of

the test taker. In some cases, the direction appeared to be reversed, with markers of higher fluency combining with nonverbal behavior to produce an impression of greater confidence. When asked what made sample 13 appear more confident, the rater 6 said:

> *[He] didn't seem to be stumbling a lot, and wasn't looking around and adjusting and fidgeting a ton. So his voice seemed to remain like, I guess not shaky. So that's why I said, you seem to be pretty confident and not anxious. Um, overall, the sentence seem to flow pretty well, too.*

The rater noted that behaviors such as shifting gaze, the use of random gestures and self-adaptors, particular features of tone, and fluency features all combined to produce a confident impression of the test taker. In another case, rater 14 mentioned that sample 29's confident affect improved once the breakdown sequence concluded:

> *So when she started rolling with her response, it was very fluent. And there weren't very many hesitations. She had some of those same grammar issues with the missing plurals and stuff. But she was very comprehensible. And then her confidence seemed to gain once the, I'm sure that, I think the person was like, "Yes," or... And she seemed like very pleasant to be there. The inflection her voice was warmer and happier.*

In this case the reverse direction from fluency to confidence was much more apparent. This comment also links back to the pattern of adaptability. The test taker's adaptable nature to the breakdown sequence led to greater confidence, and hence an overall stronger impression of language ability. These bi-directional relationships between fluency and assuredness were almost always linked to overall positive impressions of the individual.

On the other hand, observations of anxious or unconfident test takers did not frequently relate to a direct negative impression of language ability, as the two were seen as interrelated. For example, when discussing sample 21, rater 11 said:

> *Yeah, here definitely was where I establish she just wasn't, I think I literally put it like all the way down for anxious… I accept, pretty much established she was completely unconfident and anxious at this point, because she just looked like kind of, like her facial expressions and stuff. And also,*

she just kind of seemed like she just didn't understand like, what was going on? So that's that that to me, put, why I put her so far.

Here the rater noted that the test taker appeared quite anxious due to her nonverbal behavior but noted that this might have been related to her lack of comprehension. Although the overall judgement appeared negative, it did not relate directly to language proficiency. In fact, raters appeared to be able to separate affect from proficiency in at least some of these instances, such as rater 16's comment about sample 16: "He seems very unconfident. But the words that he's using, he clearly knows. Like, he clearly has a good vocabulary. He just, it just doesn't seem like he's confident enough to use it." Despite being seen as lacking confidence, the rater still found the test taker's vocabulary level to be strong. Thus, judgements of higher proficiency may be facilitated by a presentation of assuredness, but they are not necessarily hampered in the lack of such affect.

Summary

This chapter has described the patterns and trends that emerged from raters as they described their thought processes while rewatching and listening to stimuli from the main rating project. As opposed to the analysis of their scores, using stimulated verbal recall provided insight into where the raters directed their attention and what information they used when making judgements about language proficiency. The raters' comments on nonverbal behavior made up a small but substantial portion of the dataset, aligning with previous studies that used raters to elicit comments about L2 communicative competence. The raters in this study focused primarily on behaviors related to eye gaze, mouth movements, and paralinguistic cues, with a smaller focus on body posture, gesture, and eyebrow movements. The raters' comments revelated that they rarely focused on just one behavior when making decisions related to language proficiency or affect. The raters considered ensembles and the various conflicting information from both verbal and nonverbal behavior when making decisions.

The raters' comments revealed four main patterns related to how they used nonverbal behavior when rating language proficiency. The first was that they used multimodal information from both nonverbal and verbal channels to gauge whether a test taker understood their interlocutor. Non-understandings always

led to a negative impact on language ability. The second pattern, however, was that this negative impact could be attenuated by engaging in adaptable behaviors—behaviors that showed a desire to communicate, engagement, active thinking, and interactional competence. The third pattern was that comprehensibility was judged using both verbal and nonverbal criteria. Fluency, vocabulary, grammar, and pronunciation worked together with positive affect and engagement to aid in smooth communication that was easy to understand. Finally, the fourth pattern was that assuredness in the form of confidence and low anxiety was closely related to judgements of language proficiency; the more assured someone appeared, the more proficient they were perceived. This relationship was, however, bidirectional.

CHAPTER 8: DISCUSSION

In this chapter, I will discuss the findings from Chapters 5 through 7. For each chapter, I review the findings and offer an interpretation of their significance in terms of previous findings and theory. For Chapters 6 and 7, I will discuss the results of the *a priori* hypotheses from Chapter 3. These three sections will be followed by a triangulation of the three sets of findings, focusing on the themes these studies have in common. I will then discuss the implications of this study on L2 assessment, SLA research, and research methodology.

Study 1: Affect and language proficiency

Dataset integrity

Chapter 5 began by ensuring the dataset was as robust as possible, as survey and questionnaire data is notoriously prone to problematic rater responses (Iwaniec, 2019). In this study, I planned for several *a priori* measures to minimize the impact of such behavior. For example, raters were not allowed to speed through the survey, as they could not move forward through the samples without watching the videos in their entirety. Raters were also not allowed to pause or replay the videos, as this could have caused differential effects in the ratings if they focused too much on particular language elements during replay. To reduce straight lining (the selection of all categories on one side of the scale) and acquiescence (the selection of categories that a participant feels the researcher desires), I randomly reversed the polarity of scales. This encourages raters to consider the meaning of the scale for each case, and in theory reduces this type of bias. To reduce order effects of the presentation of the samples, the samples were counterbalanced by day and randomized. To reduce primacy bias in the affect scales, scale ordering was randomized for each rating sample as well.

However, these measures do not guarantee the integrity of rating data, as these data must be inspected *a posteriori* for problematic responses. I described how the dataset was cleaned to ensure the ratings were as reliable as possible prior to analysis. Of the 100 original samples, one was removed because of technical problems, and 16 were removed due to low reliability, misfit, multivariate outliers, or a combination of these issues. Although other methods exist to detect problems in survey-type data, such the

careless package in R to detect careless responding (Yentes & Wilhelm, 2018), some of these are not effective for data with a nested structure (raters by samples with multiple outcomes). For this reason, I relied on methods highlighted here, which I decided would be the most straightforward method to strengthen the dataset without losing unnecessary power.

I then checked the integrity of the dataset prior to conducting analysis using descriptive statistics and Rasch measurement. The scales functioned appropriately without misfit or erratic behavior, showing desirable Rasch measurement characteristics. Raters used the full range of scores, and the standard deviations for each rating category indicated that raters showed a satisfactory degree of variance across the 7 score categories. Language-related traits had the highest amount of variance, while affect measures, in particular positive affect, had the lowest. For this study, this type of variance was important as it suggested that raters as a whole were not assigning a restricted number of scores for the language categories, which would have attenuated correlations and weakened inferences from regression analyses. The distribution of the scale scores furthermore showed desirable characteristics. Overall, raters awarded midpoint scores of 4 less frequently than scores indicating a scale direction (e.g., 3 or 5, marking direction towards a descriptive endpoint such as positive or negative). I understood this as indicating that raters did not exhibit central tendency. Language scores were fairly balanced overall, with somewhat bimodal distributions (3 and 5 were the most frequent categories for fluency, grammar, and vocabulary; comprehensibility was negatively skewed). Anxiety and confidence were similar. Other affect scores tended to be skewed negatively, showing that raters assigned higher scores for these phenomena. All scale categories correlated at a medium to strong level according to Plonsky and Oswald (2014). These correlations suggest that there was very likely an impact of halo effect, with raters assigning similar scores across all categories, despite the fact that the scales were randomized and with random polarity. In terms of severity, the Rasch partial credit model showed that grammar was the most difficult scale, and comprehensibility was the easiest.

I also checked the integrity of the samples. The samples showed desirable Rasch statistics, with no sample showing any misfit. This was positive as it indicated that there were no problematic samples that caused large degrees of inconsistency in the ratings. The samples were selected to represent a wide range

of abilities (on the CEFR, A2–C1), and the language proficiency scores the raters awarded were largely consistent with the base proficiency scores awarded by IELTS. The scores on the samples also showed that the raters detected a wide variety of differences amongst the samples in affect measures. Curiously, as noted with the correlations above, these affect scores showed a linear tendency with ability level, especially with categories such as confidence. Other categories, such as attention and engagement, showed much less of a linear relationship with base proficiency.

The raters also largely showed desirable Rasch fit statistics. As expected, however, their consistency and agreement were limited. Exact agreement was low, and ICCs were largely low or medium. This was anticipated, given the minimal instructions and practice, and lack of benchmarked samples for rater training. In fact, I considered this degree of agreement positive given that raters were novice, and the scales were simplistic. The raters had never rated language categories before, and likely had never explicitly thought about these characteristics when listening to a L2 speaker. The fact that ICCs were near .5 suggests that there was a limited but shared understanding of some of the underlying characteristics of language. Rater training could have boosted these ICCs, but an explicit focus on language would have narrowed the raters' focus to the linguistic code. This might have removed necessary score variance from the use of information in the visual realm, and it would have distanced the participants from real world listeners.

RQ1: What is the relationship between interpersonal affect and language proficiency?

To answer this question, I first reduced the dataset using exploratory factor analysis. This resulted in four factors, which I named language, assuredness, involvement, and positivity. Assuredness was a combination of confidence and low anxiety, while involvement was a measure that represented engagement, attention, and interaction. Positivity represented positive affective measures of warmth, happiness, expressiveness, and positive attitude. Each of these factors was found to correlate with language proficiency measures (polychoric correlations .43–.72). Using factor scores, I then ran ordinal mixed-effects regression to determine which of these three factors related the most to the four language proficiency measures. By doing so, I provided evidence of differential impact of these measures on the language outcome variables.

Assuredness alone predicted changes in fluency and grammar scores with a strong effect size for

fluency, and a smaller effect size for grammar. Assuredness was also one of three significant predictors of vocabulary, with a smaller effect size. These findings indicated that as an individual is perceived as more confident and less anxious, they are also seen as having stronger fluency, grammar, and vocabulary. This finding is largely consistent with the literature. Anxiety, for example, has frequently been found to relate to lower L2 proficiency or achievement outcomes (Botes et al., 2020; Clément et al., 1980; Clément & Kruidenier, 1985; Dewaele & Alfawzan, 2018; Dewaele & Li, 2022; Dewaele & MacIntyre, 2014; Dewaele et al., 2019; Jiang & Dewaele, 2019; Jin et al., 2017; Li et al., 2020; MacIntyre et al., 1997; Teimouri et al., 2019). Confidence, on the other hand, has been found to predict educational achievement (Ahammer et al., 2019; Cobb-Clark, 2015; Heckman et al., 2006; Judge & Hurst, 2007; Stankov et al., 2012) and L2 achievement outcomes (Doqaruni, 2015; Edwards & Roger, 2015; Labrie & Clément, 1986). Confidence is an affective stance that raters frequently observe and factor into positive evaluations of test takers in the language testing literature (Jenkins & Parra, 2003; Neu, 1990; May, 2009, 2011). Given the close relationship confidence and anxiety have with cognitive, psychological, and personality elements (e.g., Stankov et al., 2012), raters may have drawn on nonverbal cues to extrapolate about the test takers' underlying cognitive fluency and lexicogrammatical competence through the affect they demonstrated. Seeing an individual as anxious may have led raters to doubt the individual's level of ability, resulting in lower scores awarded. Likewise, seeing a confident performance could inform raters that the test taker believed in their own abilities, thus leading to higher gains.

There are caveats that should be mentioned, however. Through the stimulated recall in Chapter 7, I found that most raters used a broad definition of fluency (Lennon, 1990), which they often understood as overall language ability (that is to say, "He speaks English fluently" for linguistic laypeople generally indicates that the speaker is perceived as proficient across multiple aspects of language). For this reason, it may be erroneous to think of fluency in this case in the narrow, psycholinguistic sense of an "impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently " (Lennon, 1990, p. 391), such as discrete forms of utterance fluency (e.g., pauses, repair, articulation speed; Segalowitz, 2010). The fact that assuredness impacted three similar areas

of language proficiency is perhaps evidence that confidence, anxiety, and overall communicative ability are closely bound together. It is also important to note that this is not evidence of a *causal* relationship, as confidence and proficiency likely have a bidirectional relationship (Edwards & Roger, 2015). Evidence of confidence and low anxiety may lead to an impression of greater proficiency, and at the same time, greater proficiency, and ability to handle a communicative event likely leads people to display a greater amount of confidence and ease when speaking.

Involvement, a combined measure of engagement, attention, and interactiveness, was the sole predictor for comprehensibility, with a large effect size. Involvement also predicted vocabulary scores, along with assuredness and positivity, with a smaller effect size. The findings here show that as individuals were seen as more engaged, attentive, and interactive, they were easier to understand and their vocabulary was perceived as stronger. Engagement has been defined as a person's level of interest and participation in an event (Philp & Duchesne, 2016), which may be likewise perceived as a desire to communicate. In the language testing literature, engagement has been detected by raters through mutual gaze, smiling, head nods, and a forward leaning posture (Ducasse & Brown, 2009, Jenkins & Parra, 2003; May, 2009, 2011; Nakatsuhara et al., 2021a; Neu, 1990; Sato & McNamara, 2019), all of which are also closely related to attention and interaction. Studies in SLA have furthermore found links between both subjective observations of engagement (as collaborativeness; Nagle et al., 2022) and objective measurements of head nods (Trofimovich et al., 2021) with comprehensibility. Trofimovich et al. (2021) found that different dimensions of engagement, including social involvement (e.g., using encouraging language) and backchanneling (through nodding), successfully predicted comprehensibility. Nagle et al. (2022) likewise found that comprehensibility was predicted by collaborativeness (social engagement), and low anxiety. This study supports these findings here in that involvement was a strong predictor of comprehensibility. Assuredness, with included measures of low anxiety, was not a predictor of comprehensibility in this model, although the correlations were positive between these measures. Regarding the association between involvement and vocabulary, no study to date, to my knowledge, has shown a link between these measures. It could be postulated that when individuals are more involved, they show a stronger degree of willingness

to communicate. If this link stands, there is some recent literature that supports the connection between involvement and vocabulary, namely a greater amount of productive vocabulary use (Heidari, 2019) and receptive vocabulary knowledge (Şen & Oz, 2021) in learners that are more willing to communicate.

An alternative explanation for the impact of involvement could reside within the literature on affective contagions (Elfenbein, 2014; Hatfield et al., 1994). When a rater sees a test taker as involved, it is possible that some of that involvement transfers to the rater, making them feel more invested in paying attention and listening to the test taker. In other words, willingness to communicate may inspire a *willingness to listen*. Raters that are more willing to listen to a particular speech sample would likewise be more likely to find it easier to understand, because they were already making the effort to do so. This interpretation would have important implications for research on measuring comprehensibility if purely linguistic forms were the object of study and the individual were able to be seen. It would also open the door to more literature on this topic in language testing, as the impact of interlocutors or examiners on test takers' performance outcomes has been studied, but there is a need for work considering the impact of test takers' behavior and language on rater behavior (Briegel-Jones, 2014; Brown, 2003; Plough & Bogart, 2008).

Regarding positivity (positive affect), the only significant relationship in these models was a negative relationship with vocabulary. This finding is curious, as the correlation between positivity and vocabulary was not negative, but positive, with a medium effect size. I suspect that this negative coefficient is a statistical artefact or the result of a complex relationship with assuredness and involvement, and for that reason I hesitate to interpret this effect as meaningful without more data. Past literature states that positive psychology plays an important role in language learning (MacIntyre et al., 1998; MacIntyre et al., 2019; Oxford, 2016), helping learners overcome anxiety and creating opportunities to learn (MacIntyre & Gregersen, 2012;). Several studies have also found links between language achievement and enjoyment, which may be a facet of positive affect (Botes et al., 2020; Dewaele & Alfawzan, 2018; Dewaele & Li, 2022; Dewaele & MacIntyre, 2014; Dewaele et al., 2019; Jiang & Dewaele, 2019; Li et al., 2020). Trofimovich et al. (2021) also showed that positive affect related positively (to a very small degree) with

comprehensibility gains in some of the participants, but not all. Chong and Aryadoust (2023), on the other hand, found no evidence of an impact of positive affect on language proficiency measures. Thus, the null findings in Chong and Aryadoust (2023) and the negligible effects in Trofimovich et al. (2021) align with the findings in this study that positivity was not a significant predictor of fluency, grammar, and comprehensibility. It is perhaps the case that positive affect may relate to achievement outcomes in classroom-based settings, where teachers and students interact with each other in a much more relaxed environment, but not in a high stakes situation where language is being scrutinized. More research is necessary to determine whether an effect indeed exists.

A key question at this point is whether the findings here are meaningful in a language testing context. If one considers what is construct relevant or irrelevant, it is intuitive to consider positive affect as being construct irrelevant to any measure of language. There are many reasons a highly proficient candidate may exhibit markedly lower positive affect in a testing scenario. For example, if a test taker is somewhat less expressive, it may be because of the stressful testing context. Concentrating on producing accurate language while maintaining rapport with the examiner in a social setting may spike cognitive load, inhibiting the display of particular affective stances. In addition, a broad range of neurodiverse test takers may display differing patterns of affective phenomena or differing patterns of nonverbal behavior, such as gaze differences and repetitive motion in individuals with autism (American Psychiatric Association, 2013). In these cases, language ability should not be judged worse in these individauls than someone who smiles more or maintains more mutual gaze on that basis alone. If this finding were true, test takers may find it beneficial to train their behavior in order to get a higher score. It is fortunate, then, that positivity was not a major predictor in any of the models.

The construct relevance of assuredness and involvement, however, is more complex. If indeed assuredness is a bi-directional result of language proficiency, and if assuredness is partially a cognitive mechanism, being perceived as more confident or less anxious may in fact reveal something about the person's ability to speak. I say this with caution, however, as I would not argue the opposite: Testing situations are anxiety-producing, and the mere fact of feeling anxious or having lower confidence should

not reveal anything about lower language ability. A similar interpretation is possible for involvement. Showing engagement, attention, and interactiveness are all critical skills in L2 communicative competence, especially within subdomains such as interactional competence (Galaczi & Taylor, 2018; Plough et al., 2018) or goal-directed communicative effectiveness (Morreale et al., 2013). Being engaged, interactive, and attentive are all crucial to effective communication amongst neurotypical individuals, as these stances help people build bridges in interpersonal encounters, especially when problems or breakdowns occur. Displaying this type of affect alone is not enough to overcome very low proficiency, but it can help facilitate intercultural encounters. Appearing withdrawn, disengaged, and inattentive can have an opposite effect, causing communication to deteriorate or fail.

Neurodiverse individuals, however, may not exhibit the same range of attention and engagement as neurotypical individuals, especially in a testing context. Neurodiverse test takers, such as those on the autism spectrum, may exhibit full communicative competence within a different set of behavioral repertoires. To my knowledge, it is unknown how raters interpret these stances in L2 settings, and more research is needed in this area. Raters would need to be trained on how to work with these test takers, as particular accommodations may be necessary to ensure a bias-free, equitable speaking test (Randez & Cornell, in press). In any case, overall, raters naturally pick up on both assuredness and involvement, as attested in many studies, and for this reason it is likely a natural part of the construct of speaking, but any operationalizations of these types of affect would have to be carefully implemented, and accommodations for various populations would have to be thoroughly researched.

Study 2: Automated measures of nonverbal behavior

Measurement patterns

In Chapter 6, analyses of the iMotions behavioral output showed substantial variance in the nonverbal behavior of the samples. The samples varied the most in engagement (in the iMotions data, a measure of expressiveness, or amplitude of facial muscle activation), both in the mean and standard deviation values. This showed a spread of different profiles in overall facial movement and how much the facial movement varied within each sample. The variance in this measure indicates that as test takers

completed the test, they expressed a range of different visible facial movements, ranging from more neutral to highly active. Note, however, that the measure of engagement was not a measure of whether the expressions indicated any certain positive or negative direction.

Measures of valence indicated whether expressions were positive or negative, yet this measure showed very little overall variance. Most test takers maintained a fairly neutral stance, and the time series graphs showed generally stable patterns for most test takers within each sample. This is not altogether unexpected, however, as it is possible that the institutional nature of the testing context encourages some test takers to manage impressions of how they are perceived (Luk, 2010), which may then restrict strong expressions of positive or negative emotions. Culture may have also played a role in these differences, as it is possible that Chinese individuals may manifest a different set of display rules regarding their emotional expression than their American counterparts (e.g., Ekman & Friesen, 1969; Matsumoto & Hwang, 2012). Nonetheless, there were individuals that displayed visibly positive valence during the tests. Samples 14 and 15, for example, smiled regularly throughout and had high mean valence scores. Contrastingly, few test takers showed clear negative emotions during their samples. Although the test taker in sample 8 did smile at times, he struggled with the test questions and appeared to express a more worried or concerned look during the test, which resulted in his valence score being the lowest of the group.

In terms of attention, the test takers generally had high mean attention measurements, with most scores near the maximum. Only seven samples showed values less than 90 (of a maximum of 100). This suggests that the test takers in these samples largely directed their gaze and head turns towards the camera. The time course graphs showed test takers that scored lower on attention generally varied their attention more; in other words, these individuals tended to withdraw and reestablish attention by looking at or turning towards the camera more frequently. For example, samples 20 and 24 broke their attention regularly throughout and scored the lowest on attention. Lower attention with high variance is quite common, however, with speakers in interactional settings. Generally, listeners maintain attention with their interlocutors, and speakers may look back and forth at their listener while speaking, using direct and averted gaze to manage conversational structure (Rossano, 2012; Goodwin, 1980). Thus, lower attention scores are

not necessarily a negative characteristic, given that these samples showed the test takers speaking for the majority of the duration.

It is also important to note the differences between these three measures and similar measures of assuredness, involvement, and positivity as measured by the raters in Chapter 5. The raters' subjective observations of behavior correlated quite strongly with most other observed measures, including language proficiency. Although the factor analysis was able to identify unique patterns in these measures, there is a strong argument to be made that raters exhibited a halo effect across the scales. The correlations between these factor scores and the iMotions variables, however, showed that there were communalities in what was being measured, but these were somewhat unexpected. Positivity showed a medium correlation with valence, which is logical, but it correlated even more strongly with engagement (facial muscle activation). Given that the positivity factor included the measured variable expressiveness, this may be the reason for the medium correlation with both measures. I consider these correlations evidence that iMotions data were indeed measuring aspects of nonverbal behavior. The medium rather than strong correlations (for example, between positivity and valence) may be a result of differences in what is perceived and what is measured, as there is evidence that externally measured and rater-perceived aspects of behavior do not always align (Gullberg, 1998). Given that the positivity variables (warmth, happiness, attitude) were quite subjective for most raters, a medium correlation with iMotions variables is positive for this study. I also believe this makes a strong argument for the use of objective measures of behavior here, as these were measured independently of language, unlike the rater variables.

RQ2.1: Do externally measured indices of nonverbal behavior predict language proficiency scores?

H2.1.1: Indices of attention and expressiveness will in significant but moderate correlations with language ability.

H2.1.2: Higher values of attention and expressiveness will result in significant positive regression coefficients of fixed effects, indicating an overall positive impact on impressions of second language proficiency across ability levels.

There were two broad patterns that characterized the relationships between the iMotions predictor behavioral means and the language proficiency outcomes. The first involves fluency, vocabulary, and grammar ratings. Each of the sets of predictors behaved quite similarly with these three outcome variables. In each case, the only behavior that correlated significantly with each outcome was valence, with correlations ranging from .07 (grammar) to .12 (fluency). These correlations suggest a positive relationship between overall positive expressions of emotion and how the raters perceived language ability. Correlations with engagement (.02–.04) and attention (-.02– -.01) were low but non-significant. These findings align somewhat with the literature, as test takers that appear more open, outgoing, and expressive (perhaps through positive expressions) in the testing situation may be seen as more proficient overall (Jenkins & Parra, 2003; May, 2011; Neu, 1990) and display greater fluency (Kim et al., 2023; Tsunemoto et al., 2022), though whether this is a direct reflection of proficiency is a question still up for debate.

There was a second pattern that emerged regarding comprehensibility. Valence and engagement both emerged as significant positive correlations for comprehensibility (.18 and .11, respectively), but attention was not (-.03). These correlations suggest that both the amount of expressiveness *and* its positive direction related to greater comprehensibility. In other words, more expressive or more positively emotive test takers were perceived as easier to understand. These correlations may also reflect a willingness to listen to the samples on the part of the raters, as positive affect may have encouraged raters to pay more attention to what the test takers were saying. In both of these cases, it is possible that the raters saw positive valence and expressiveness as a sign of confidence or even task engagement, of which positive affect has been theorized as a component (Philp & Duchesne, 2016; Trofimovich et al., 2021). This engagement may have led the examieners themselves to be more engaged, thus leading to higher comprehensibility scores.

The fact that attention did not correlate with any of the proficiency outcomes (-.03– -.01), was unexpected. Studies have repeatedly found that mutual gaze relates to positive impressions of test takers, while averted gaze exerts a negative influence (Choi, 2022; Ducasse & Brown, 2009; Jenkins & Parra, 2003; May, 2009, 2011; Nakatsuhara et al., 2021a; Sato & McNamara, 2019). I can hypothesize two possible reasons for this lack of correlation in this study. One, the iMotions index of attention is not restricted to eye

gaze behavior alone, as it also includes head turns. Thus, this measure was not a sole measure of gaze, which might have been more informative of attention in this case. Second, it is possible that mean attention is not informative as a measurement itself. Rather, the frequency of changes in gaze patterns or the way test takers break and reestablish gaze to convey interactional information depending on the underlying social context of interaction may instead be more informative for raters. Both of these issues will be explored later in this discussion.

The correlational analysis thus partly confirmed hypothesis 2.1.1. Expressiveness (iMotions engagement) correlated with comprehensibility, as hypothesized, but the correlation was quite small (.11). Correlations with attention and all outcomes (-.03– -.01), on the other hand, were non-significant, negating this aspect of the hypothesis. Not hypothesized was the role that valence would play in the correlations; valence exerted the strongest effect on all measures (.7–.18), with comprehensibility being the highest (.18).

In the regression analyses, however, the main effects of the mean behavioral indices did not emerge as significant predictors of the outcomes. This indicates that despite positive correlations, the main effects did not predict language proficiency outcomes across ability levels. Only interaction terms in the interaction models showed significance. Thus, hypothesis 2.1.2, that attention and expressiveness would predict proficiency outcomes across all test takers, was refuted. This finding did not support studies such as Kim et al. (2023) and Tsunemoto et al. (2022), in which a greater number of smiles (positive valence) and eyebrow behavior (expressiveness) enhanced fluency scores, and averted gaze positively predicted comprehensibility.

RQ2.2: Do nonverbal behaviors impact outcomes differentially depending on the base proficiency levels of test takers?

H2.2: Significant interaction coefficients of base language proficiency with attention and base proficiency with expressiveness will indicate that the effect of nonverbal behavior on rated outcomes depends on the base proficiency of the test taker.

For fluency, grammar, and vocabulary, the best fitting models were the ones with interaction terms with base proficiency (the scaled IELTS scores) included. None of the interaction terms, however, emerged

as significant, with only the main effect of base proficiency emerging as a significant effect. The fact that base proficiency was a main effect predicted the final outcomes is positive, as it provides evidence that the raters in this study awarded scores largely in line with those of the initial proficiency assessment. This suggests that even though the raters were novice and untrained, they shared a common understanding of language proficiency with the underlying construct of the test. The non-significant interaction coefficients indicated that for fluency, grammar, and vocabulary, even though valence correlated with the outcome, it did not emerge as a significant predictor of the raters' scores, and test takers with varying proficiency levels were not impacted differentially. This finding is somewhat surprising, as the literature suggests that less proficient test takers may benefit from being more expressive or positive than stronger test takers (Jenkins & Parra, 2003). It may be the case, though, that the effects of overall behavior on judgements of language proficiency are generally so small that they are rendered non-significant by the overall effect of language.

Comprehensibility showed a different trend, however. The best fitting model was again the interaction model, but in contrast with the models of fluency, vocabulary, and grammar, there was a significant interaction between base proficiency and valence. In order to explore this relationship graphically and also through statistical approaches, I dichotomized base proficiency into low and high groups by using a B2 cutoff score. Results showed that lower proficiency (< B2) test takers benefitted from positive valence in their comprehensibility ratings, and there was little change in the higher proficiency group (≥ B2). This aligns well with Jenkins and Parra (2003), as raters found the test takers easier to communicate with as they were more generally expressive, perhaps through positive affect. This ease of understanding in the original study may have caused raters to have a more positive impression of the test takers' overall communicative ability, thus raising scores. These findings may also align somewhat with Trofimovich et al. (2021), who found a partial positive correlation of .12 between positive affective behavior and comprehensibility in one group of dyads.

Somewhat more difficult to explain is the apparent negative impact of higher valence on the more proficient test takers' comprehensibility ratings. However, the range of valence measures in the more proficient group was markedly more restricted, which may have attenuated correlations or even muddled

inferences as a statistical artifact. It is likely the case that more proficient speakers are perceived as more comprehensible regardless of their behavior, and it is unlikely that increased valence would relate to a more proficient person being less comprehensible. This can only be tested if a greater range of behaviors is measured in the two groups.

Relationship between variance in behavior and language proficiency

After drafting and preregistering the study (but prior to running the analyses) (Burton, 2021b), I decided to analyze the previous research questions with predictors of standard deviations of the iMotions variables. I wanted to test whether behavioral variance instead of mean values could exert an effect on the model. Thus, I ran the same models and analyses with these alternative indices. Similar to the models using mean predictors, the exploratory models using standard deviations also exhibited two patterns, one impacting fluency, vocabulary, and grammar scores, and the other with comprehensibility measurements. Each fluency, vocabulary, and grammar outcome were impacted similarly by the variance of the predictors, with small negative correlations with the standard deviation of valence, small positive correlations with the standard deviation of attention, and non-significant correlations with engagement. These correlations indicated a relationship between shifting between emotive states (positive, negative, neutral) and lower scores, while shifting attentional patterns corresponded with higher scores.

Regression modeling of these effects showed that, again, the model with interaction effects best fit the data. As opposed to the average predictors, however, the interaction term between attention and base proficiency (the scaled IELTS scores) was significant for fluency, vocabulary, and grammar. Only in fluency was the main effect of the standard deviation of attention significant. Other interactions and main effects were not significant. These findings suggest that the impact of varied attention differed according to base proficiency level. In a follow up analysis, I found that less proficient speakers tended to be negatively impacted by variance in attention, while more proficient speakers were positively impacted by this variance.

While this finding may not be immediately intuitive, it does align with findings from past research. Gaze, a major component of attentional focus, is a complex behavior that may change according to various

cognitive states, social cues, and pragmatic needs. Shifting between mutual and averted gaze is an important aspect of interactional moves in speech, with speakers showing uptake (that is, integration) of information and initiation of turns with the breaking of gaze, and gaze returning to the interactant when turns are complete (Rossano, 2012; Goodwin, 1980). Speakers may also return gaze to their interactant to create a gaze window, offering listeners the chance to backchannel to show intersubjectivity (Bavelas et al., 2002). Speakers may also break gaze when questions are more difficult (and perhaps not understood) as a way to wrangle additional cognitive resources (Burton, 2023; Doherty-Sneddon & Phelps, 2005). In language testing contexts, less proficient speakers have been found to use ensembles of behavior (in particular, gestures) that are more frequent, more irrelevant to the content of their utterances, and narrower in range, while more proficient speakers may use more integrated verbal-nonverbal utterances (Gan & Davison, 2011). Irrelevant, non-target-like gaze patterns may also become salient and informative to raters. Together, the attention data in this study tell essentially the same story. Speakers that varied their attention in an integrated way with their utterances were more likely to be perceived as producing more fluid speech with stronger vocabulary and more accurate grammar. Attention, then, added to their overall impression of language ability, as they were able to use attention as a tool at their disposal to manage the interaction with the examiner. On the other hand, when less proficient speakers varied their attention more, it may have signaled that the test takers were struggling to cope with the topics or questions. These attentional shifts would not have been integrated with speech as with the more proficient group, and the raters may have used these shifts as evidence for weaker language ability. This largely goes against previous findings that averted gaze has a negative impact on raters (Choi, 2022; Ducasse & Brown, 2009; Jenkins & Parra, 2003; May, 2009, 2011; Nakatsuhara et al., 2021a; Sato & McNamara, 2019). These findings suggest that breaks in attention have a far more complex relationship with L2 speaking test scores than hypothesized.

For comprehensibility, patterns with the standard deviations were markedly different. While attention also showed a small positive correlation (.11), valence did not (-.03). Engagement in this case, as opposed to the previous model, had a small positive correlation (.06). Nonetheless, while the interaction model was again the model that fit the data best, none of the predictors explained the variance in the model.

This indicates that varying attention did not have the same impact on comprehensibility as with fluency, vocabulary, and grammar, despite the positive correlation with this measure. Given that mean valence impacted comprehensibility but varying attention did not, it may be the case that comprehensibility may be a construct more impacted by positive affect and engagement. This would align well with the findings of Nagle et al. (2022), who found that lower anxiety and greater collaborativeness served as predictors of comprehensibility scores. Likewise, Trofimovich et al. (2021) found that behaviors such as nodding, which is evidence of interactional engagement, related to positive outcomes in comprehensibility (r = .03–.34 in contrasting sets of pairs), with some indication that smiles may have played a role as well (r = -.28–.12 in contrasting sets of pairs). Tsunemoto et al. (2022), on the other hand, found that looking away predicted comprehensibility, which may have been a sign of thinking about content and relating to engagement, and would appear to align well with the positive correlation above. Likewise, they found that eyebrow movement predicted lower accentedness. Nonetheless, these authors did not provide standardized effect sizes for comparison across studies. *Positive* affect, task collaborativeness and engagement, and lower anxiety may then stimulate a willingness to listen in the interactant, which would undoubtedly cause test takers' to be more easily understood.

The findings from study 2 are overall positive. The effects found, though significant, were all quite small and explained very small amounts of variance in the models (2–3%). It is unlikely then that attention, valence, or engagement would cause large score differences alone. While many aspects of nonverbal behavior are construct-relevant aspects of communication, behavior probably plays much less of a role in conveying grammar and vocabulary ability, though it is certainly relevant with certain gestural forms that encode, for example, path motion (Kita & Özyürek, 2003; Slobin, 1996, 2006; Talmy, 2000). Growing evidence, especially from the past two studies, shows that nonverbal behavior may be highly useful for raters when understanding aspects of fluency (e.g., sources of breakdowns) or being able to comprehend someone well. The features that play the greatest role in impacting these proficiency outcomes need to be outlined, as any inclusion in a speaking test construct would need evidence that indeed they reveal

something about language development. The raters in the stimulated recall study helped to triangulate the findings of the first two studies, providing additional evidence towards this broad goal.

Study 3: Stimulated Verbal Recall

The stimulated recall sessions conducted in Chapter 7 served as the explanatory sequential mixed methods (Creswell & Plano Clark, 2017) component of the dissertation. It was explanatory in that it served to explain further the reasons for score changes in the larger dataset by highlighting raters' thought processes. It was sequential because the MFRM analysis served to identify samples that deviated from the base proficiency ratings the most. Twenty raters took part in this part of the study, and each rater watched 10 videos that they had previously scored online. These 200 recall sessions revealed patterns related to the two research questions pertaining to this chapter.

In this analysis, I was first interested in validating the inferences from the stimulated recall by investigating whether the raters were aware of the general aims of the study. This was an important point, because some of the survey questions at the end of the rating study contained some, albeit minimal, reference to nonverbal behavior. Also, the raters could have investigated my research interests online, drawing their own conclusions about the research questions underlying this study. Fortunately, the categories the raters commented on did not skew towards nonverbal behavior. Raters commented most frequently on aspects of language, which was anticipated as the rating study primarily focused on language, and the language scales were first in order. After language, the raters focused on aspects of affect. This is not common in past studies on rater cognition (e.g., May, 2011; Sato & McNamara, 2019), but explicit inclusion of rating categories of affect drove the raters' focus to this area. Test interaction, including a focus on the content of utterances, was the third most frequent category, followed by a much smaller focus on nonverbal behavior.

The fact that nonverbal behavior made up 11% of the comments suggests that nonverbal behavior makes up an important part of people's decision-making processes. Similar amounts of focus on the visible realm during rating have also been found in speaking tests by Sato and McNamara (2019) and May (2011). The fact that these studies converge in the percentage of focus on behavior when the target of the recall is

language suggests that 1) raters indeed orient to the target construct and focus on language itself, but 2) they reinforce their decisions using all evidence available to them, including the behavior of test takers. I considered the fact that this percentage aligned with May (2011) and Sato and McNamara (2019) evidence that any focus on nonverbal behavior was natural and not influenced by knowing about the study previously. Furthermore, an analysis of the percentages of coded comments by raters showed idiosyncratic but acceptable trends that largely aligned with the results from the group as a whole. Raters' reliance on a range of behaviors, including gesture, facial expressions, and other postural and head movements, may be due to an implicit understanding that language proficiency and development can only be determined using multiple lines of verbal and nonverbal evidence (Stam, 2006).

RQ3.1: Which nonverbal behaviors are most salient and informative to raters when scoring?

H3.1: Gaze aversion, eyebrow raises, smiling, head tilts, and inexpressiveness will be mentioned more times by raters as noted by higher relative frequencies of comments. Gesture and posture will be mentioned fewer times due to the online format of the speech stimuli.

The raters in this study observed a wide range of behaviors across the speaking samples during the stimulated recall sessions. Some behaviors were mentioned quite frequently across multiple individuals, while others were noted only once. The most frequently and extensively discussed behaviors were gaze and mouth movements. This finding aligns with Lansing and McConkie (2003), who used eye-tracking to identify where participants directed their attention in video-based samples. They found that viewers looked most frequently at the speaker's eyes and shifted to looking at the mouth area when comprehensibility was lower (such as with unclear audio or with L2 speech). This finding also aligns somewhat with those of Batty (2021) and Suvorov (2018), who analyzed the attentional focus of L2 test takers during video-based listening tests. Batty (2021) found that the examinees' focus was "mostly on the face of whomever is speaking, with only small departures from this to directly look at gestures, objects, the setting, and so on. Participants appeared to largely split their time between watching the speaker's eyes or mouth" (p. 527). Likewise, Suvorov (2018) noted that the vast majority of test takers watched the "speaker's mouth, face,

head, hands, [and] eyes" (p. 150).

The most frequently mentioned behaviors in this study were those of the eyes. Raters in this study used gaze behaviors, such as shifting gaze and averted gaze, to understand test takers' cognitive processes, in particular listening comprehension and thinking. When the examinees' eyes were marked by movement, such as eyes darting around, this was perceived as a sign of struggle either to produce language or to comprehend the test question. This also led to the attributional judgement that the test taker was anxious and unconfident. Averted gaze was also used to understand similar processes. Thus, many of the negative features of averted gaze or shifting direction aligned with past research showing that these behaviors were perceived negatively (Choi, 2022; Ducasse & Brown, 2009; Jenkins & Parra, 2003; May, 2009, 2011; Nakatsuhara et al., 2021a; Sato & McNamara, 2019). Nonetheless, this was not always the case. Averted gaze was sometimes seen as positive if the examinee was perceived as having understood the question and was simply preparing the content of their speech. Thus, gaze away from the rater was often used as critical evidence of fluency judgements, as raters pieced together an understanding of whether the examinee was struggling or preparing utterance content. This finding somewhat aligns with Tsunemoto et al. (2022), who found that averted gaze was a predictor of comprehensibility judgements. Overall, these gaze findings are an important deviation from past research, as they suggest that context mediates the impact of averted gaze. Mutual gaze (when it was not staring), while also possibly used as a marker of comprehension, was often discussed as it related to engagement and confidence. Especially during repair sequences, when test takers maintained mutual gaze with the camera/interlocutor, this behavior led raters to understand that the examinee was making an effort to communicate and show attention, somewhat ameliorating the negative impact of the comprehension breakdowns.

Mouth movements were discussed almost entirely within the context of the test taker's affective state, aligning with Batty (2021) and Coniam (2001) in that individuals largely used facial expressions to determine attitudes and affect. The most frequent of these behaviors was smiling, which led to examinees being perceived as happy, positive, warm, and often confident and at ease. Likewise, observations about a lack of smiling or lip biting were often associated with anxiety or being unconfident. Mouth movements

did not appear to relate directly to any judgements of fluency, vocabulary, or grammar, but they did appear to have an impact, or perhaps an indirect impact, on comprehensibility. Apart from one case of an individual that was mumbling, which led a rater to try to read the examinee's lips, raters did not discuss lip reading as a tool to enhance comprehensibility, although this has been documented with L2 speakers listening to others (Hardison, 2018; Inceoglu, 2016; Lansing & McConkie, 2003; Suvorov, 2018). Instead, perceptions of positive affect, often derived from smiling and other behaviors, created a sense of approachability and led raters to *want* to listen more carefully and comprehend the test taker, thus enhancing the test taker's overall comprehensibility.

Paralinguistic features also made up a sizable portion of the comments in this dataset, relating to both fluency and overall affect. Raters frequently discussed filled pauses and slow speed as evidence of lower fluency, both of which are well documented in the literature (Suzuki et al., 2021). In some cases, filled pauses were also mentioned as a technique for pausing to think, or even as evidence of anxiety. However, most comments relating paralinguistic features to affective states dealt with laughing and prosodic features such as tone. Raters frequently used these aspects of paralinguistics in light of positive affect, which likewise added to comprehensibility.

Posture was a feature mentioned by most but not all examiners, and rarely coincided with judgements of language proficiency. Posture was largely linked to affect, as past research has attested (Coulson, 2004; Dael et al., 2012). Some postural behaviors, such as rocking, shaking, or leaning back, were seen as evidence of anxiety and a lack of confidence. These postures did not always lead to a negative impression of language ability but were sometimes seen as distracting, and overall negative. Leaning back in at least one case was seen as a sign of confidence and being at ease, as was attested in Neu (1990). Leaning forward, an important postural behavior in this category, was largely seen as positive. This behavior was seen as an attempt to remain engaged with the examiner and to show attentiveness and interactiveness, especially in cases during breakdowns in comprehension. In Jenkins and Parra (2003), leaning forward was perceived similarly as a sign of engagement and listening comprehension. Forward leaning may also indicate rapport with an interlocutor, presence, and involvement, while backward leaning

can convey a lack of presence, distance, and detachment (Burgoon et al., 1984; Mehrabian & Williams, 1969). Although movement in general was seen as positive, quick, erratic movements (especially with the hands) was associated with struggling and breakdowns in fluency. A rigid, unmoving posture was generally seen as negative, aligning with Gan and Davison (2011), May (2009, 2011), and Neu (1990).

Gestures made up only a small portion of the overall comments. This was at least partly because of the limited range of motion visible in Zoom recordings, as the laptop was placed quite close to the test taker as seen in the cartoonized images in Chapter 4. Nonetheless, although infrequently mentioned, self-adaptors (e.g., head scratching, adjusting glasses) were highly salient to the raters and were highly indicative of problems with underlying cognitive fluency (Segaliwitz, 2010) or anxiety. The relationship between self-adaptors and coping mechanisms for stressful situations is attested in the psychology literature (Ekman & Friesen, 1974; Kikuchi & Noriuchi, 2019) and literature with L2 speakers (Gregersen, 2005; Lindberg, 2021, 2022). These non-target-like, erratic gestures have also been mentioned as relating to having negative impressions on raters in the testing literature (Gan & Davison, 2011; Sato & McNamara, 2019; Thompson, 2016). Other gestures, such as iconic or beat representational gestures, were infrequent and isolated to one or two samples, but were unanimously seen as positive.

Head movements were also mentioned relatively infrequently, but head nods stood out as the most frequently mentioned behavior. These were nearly unanimously seen as positive. Head nods showed raters that test takers were engaged with the test discourse, following raters' questions, and showing active listening comprehension. Head nods were then a critical aspect of interactional competence, allowing the test takers to show uptake, hold the floor, and close their turns effectively with the raters. The positive impact of nodding is well-attested in the L2 literature (Jenkins & Parra, 2003; May, 2009, 2011; Neu, 1990; Trofimovich et al., 2021), and it is an important mechanism for displaying a range of information (continuers, backchannels, etc.) in online teleconferencing (Mark et al., 2023). The only case in which a head nod was perceived negatively was when it was too fast and showed a degree of anxiety.

Eyebrow movements, head tilts, and an overall lack of expressiveness were all mentioned in this dataset, but by fewer than half the raters. A whole range of other small behaviors, such as visible swallowing

and shoulder shrugging were also mentioned, but in isolated cases. These behaviors did not make up a sizeable portion of the comments, and it is difficult to generalize about their function from a small number of comments.

Thus, hypothesis 3 was only partially supported. Gaze behavior, not just gaze aversion, and mouth movements were indeed the most frequent behaviors commented on. Likewise, as hypothesized, gesture did not feature in a large number of comments, partially due to limited visibility in the online format. However, eyebrow movements, head tilts, and inexpressiveness were not frequent comments in this dataset as hypothesized. Likewise, posture was hypothesized as being a less commonly commented element, but in reality, it was mentioned by most raters. Not hypothesized was the overall importance of paralinguistic features or a general focus on the face.

RQ3.2: Relationship between nonverbal behavior and language proficiency

Nonverbal behaviors in this study were found to exert varying influences on evaluations of speakers. In general, these related to the interpretation of affect, which is in line with past research (Batty, 2021; Coniam, 2001; Kappas et al., 2013; Richmond & McCroskey, 2004; Sato & McNamara, 2019; Singelis, 1994). Raters also used nonverbal behavior to understand semantic aspects of speech (e.g., head nods and representational and beat gestures for emphasis and additional meaning), cognitive aspects (gaze shifting indicating listening comprehension and speech processing), and interactional moves (e.g., mutual gaze, head nods, and paralinguistics to mark turn-taking and holding the floor). Ultimately, these were used with speech to better understand language proficiency outcomes, but they were not used in isolation. Raters always made holistic judgements using the entire nonverbal ensemble and speech as evidence. There was never a case, for example, when someone evaluated fluency lower because of and only because of sustained averted gaze.

Comprehension and adaptability. Various patterns arose in the qualitative data that suggested that certain ensembles of behavior were useful to raters when evaluating proficiency. The most extensive and frequently mentioned of these was a multimodal assessment of listening comprehension. The raters drew primarily from nonverbal behavior, often prior to the answering of a test question, to infer about the

cognitive processes test takers were undergoing. The reliance on nonverbal cues to convey backchanneling and receipt tokens may be the result of testing in a Zoom-based format, as these exchanges have been shown to elicit far fewer verbal backchannels than face-to-face conversation (Mark et al., 2023). The raters thereafter often used this information in their judgements of overall language ability, and occasionally as it related to vocabulary or fluency when breakdowns occurred. The raters frequently described these judgements as ones of competence, one of the scale categories that closely related to language in Chapter 5. Competence was often described as both the ability to understand and successfully realize task demands. Neither listening comprehension or task success are generally aspects of rating scales (e.g., IELTS), but they regularly appear in the literature connected to raters' views of communicative competence (Brown et al., 2005; Ducasse & Brown, 2009; Orr, 2002; May, 2011; Sato & McNamara, 2019).

If raters regularly find that comprehension and task completion are part of the construct of communicative competence, this could indicate areas of construct underrepresentation in speaking tests. This could also indicate that speaking should be treated more as an integrated skill than it has to date. Given the recent resurgence of research on detecting nonunderstanding in interpersonal L2 encounters, especially from a multimodal perspective (McDonough et al., 2019, 2022b, 2023), there is growing evidence of the visual signature of non-understanding. McDonough et al. (2023), for example, found examples of facial expressions that were largely used to determine affect that aligned with episodes of understanding (e.g., engagement, attention, confidence, relaxation) or nonunderstanding (e.g., inexpressiveness, disinterest, confusion, stress). The raters in their study also discussed specific behaviors that related to these episodes. They found that gaze behavior (e.g., blank stare, looking off into space), eyebrow movements (e.g., raised), posture (e.g., slouching), and the use of self-adaptors were related to nonunderstanding, while mutual gaze and the use of representational gestures related to understanding. These behaviors and affective stances largely align with the behaviors discussed in this dissertation. With careful operationalization, one could imagine richer descriptions of listening comprehension added to rating scales for speaking tests.

The raters in this study noted that the impact of comprehension breakdowns was consistently negative. Nonetheless, that impression could be moderated by the test taker's adaptability when managing

breakdown sequences. Adaptability was marked by both affective responses and nonverbal behaviors. In terms of affect, adaptable candidates remained engaged, attentive, confident, and willing to communicate throughout the breakdown sequence. They used multimodal backchannels (e.g., nodding), natural gaze patterns showing attentiveness (either towards the interlocutor or towards thinking about content, but not specifically mutual gaze), and postural stances that conveyed engagement, such as leaning forward. They deployed multimodal resources of interactional competence to engage in active listening and to repair their breakdowns. On the other hand, unadaptable test takers were disengaged, less attentive, less confident, and appeared less willing to communicate. Their behaviors tended to be rather expressionless and distant. Their gaze was removed or shifting, they nodded very little, and showed more of a slouching posture. They also displayed fewer resources of interactional competence to manage the breakdown sequences, such as merely repeating the trouble source (e.g., "Ceremonies?") or using an open class initiator (e.g., "What?") rather than asking a clarification question (e.g., "What does ceremony mean?"). When raters noted that someone showed tendences of being more adaptable to breakdowns in comprehension, they were more likely to have an overall positive evaluation of the test taker. When fewer adaptable sequences were shown, the evaluations remained negative. In other words, test takers could overcome the negative impact of breakdowns by deploying a range of behaviors, affect, and strategies to create a more positive impression of their language abilities. This attenuating nature of adaptability could perhaps explain the relatively weaker relationship between repair fluency and perceived L2 fluency when compared with other linguistic aspects of fluency, such as articulation rate and pausing, in recent meta-analyses on fluency (Saito et al., 2018; Suzuki & Kormos, 2020).

The adaptability shown by test takers successfully managing breakdown sequences has a relevant connection to Hymes' (1972) concept ability for use. In Hymes' model, ability for use related to a person's linguistic competences that enabled L2 communication and also the cognitive, social, and affective capacity to apply their communicative skills to differing contexts. For Hymes, communicative success was then mediated by other psychosocial elements the test taker employed, which would change depending on the characteristics and needs of each situational context. A similar sentiment was echoed by Morrow (1979)

regarding interactional success: “The apparently trivial observation that the development of an interaction is unpredictable is in fact extremely significant for the language user. The processing of unpredictable data in real time is a vital aspect of using a language” (p. 16). When the test takers in these samples encountered unfamiliar or otherwise problematic utterances, they were encountering an unpredictable situation. Their adaptability to that unfamiliarity is what enabled more or less successful interactions with the examiners. Harding (2014) argued that adaptability in the essence of Hymes and Morrow might form a missing core element in models of communicative competence:

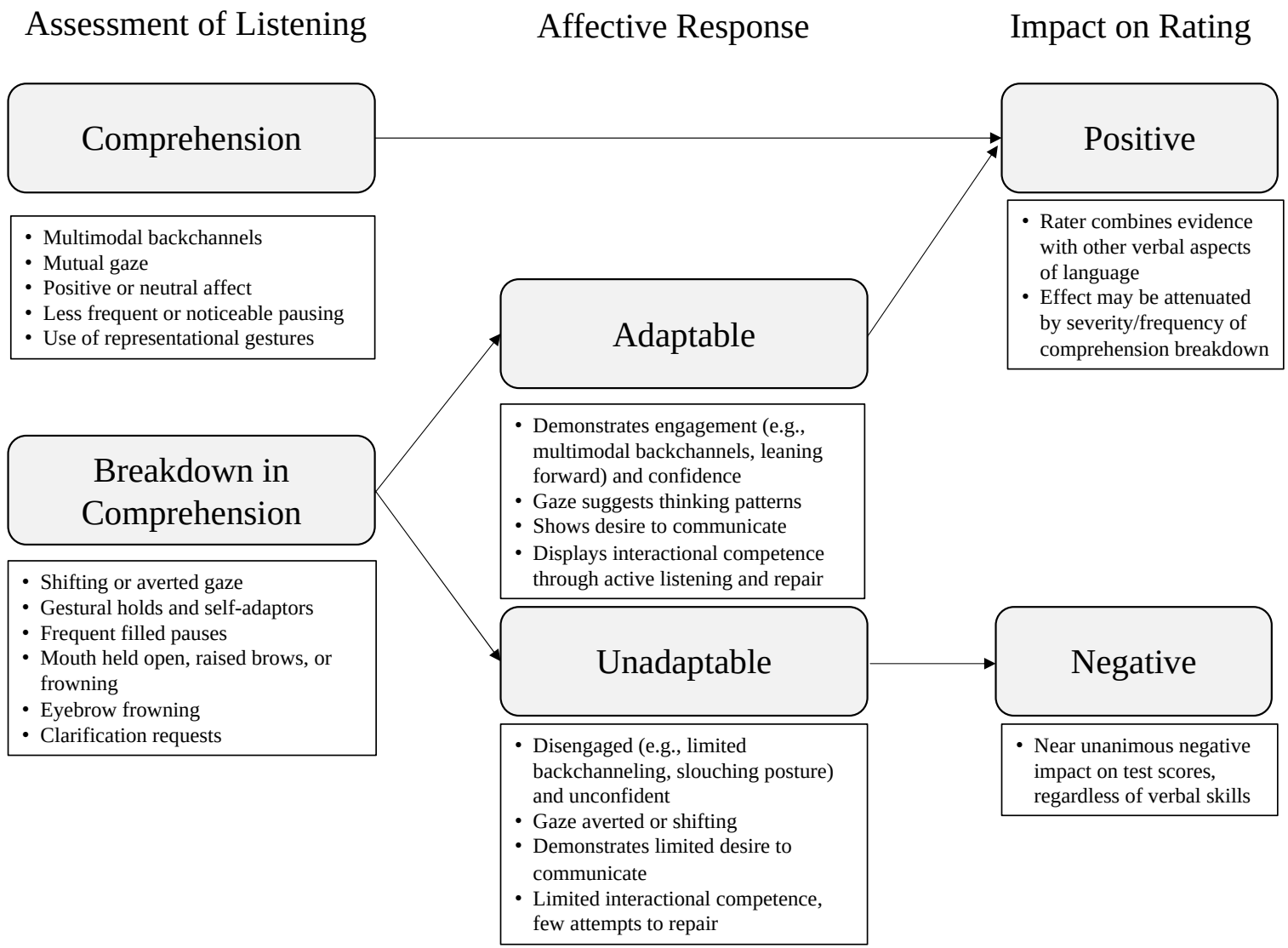
> The notion of adaptability is the common denominator in a test taker’s need to deal with different varieties of English, to use and understand appropriate pragmatics, to cope with the fluid communication practices of digital environments, and to notice and adapt to the formulaic linguistic patterns associated with different domains of language use (and the need to move between these domains with ease). (Harding, 2014, p. 194)

In line with these arguments, the raters showed an implicit understanding of the importance of adapting to unpredictable and new contexts and considered this as evidence of communicative competence when scoring samples with breakdowns.

An illustration of how I have theorized the relationship between listening comprehension and adaptability is presented in Figure 8.1. In this model, there are two unseen individuals: a rater and a test taker, or an interlocutor and a speaker. This interaction represents, in essence, the end result of a hypothetical adjacency pair: an interlocutor produces a statement or asks a question, and the test taker or speaker is expected to respond. The model begins once the rater or interlocutor hears the speaker’s response. First, the rater forms an initial assessment of listening comprehension using multimodal resources (column 1). The resources raters use listed in the model have been drawn from the literature and the findings from the raters in this study. Simultaneously, raters observe the adaptability of the test taker, with the features of adaptability being drawn from the rater reports in this study (column 2). When there are no comprehension difficulties, there is no general impact of adaptability, and the proficiency outcome may be positive if there is a display of linguistic competence in the test sample as well (column 3). When comprehension problems

surface, however, adaptability moderates the outcome impression of language proficiency. The outcome can be positive if the test taker displays an adaptable affect using pragmalinguistic tools (elements of language that convey social meaning), verbal and nonverbal interactional skills to manage the conversation, and nonverbal behavior to show an engaged and attentive affect. The outcome is negative when after the breakdown, there is no adaptability present: the test taker does not repair the breakdown and displays behavior showing disengagement.

Figure 8.1
Impact of Affect on Assessment of Listening Comprehension



The raters' intuitions about dealing with unpredictability using adaptable affective stances appears to have some precedent in the literature. The boundary between B1 and B2 on the CEFR is mainly characterized by unpredictable language use situations in which learners are able to communicate effectively (Council of Europe, 2020). The skills involved in being adaptable, which are partly pragmalinguistic in nature and partly nonverbal/affective, may be an implicit signal to raters that a speaker has acquired certain high-level skills with the language to cope with unpredictable situations, even though their core psycholinguistic structural and cognitive competences resulted in breakdowns in communication. Perhaps, then, raters prioritize these pragmatic and affective displays of competence over linguistic displays. Roever (2021) offered some support for this argument:

> High-proficiency learners at B2 and above need a strong focus on pragmatics as they have the linguistic tools needed for successful communication in a wide variety of settings, but they do not necessarily know how to deploy these tools for optimum effect… In face-to-face communication, [problems in written communication] can be compensated through tone of voice and facial expression.

If an individual *does* know how to deploy their pragmatic and affective tools for optimum effect, but lack the linguistic tools in a certain situation, this could perhaps explain the positive effect of adaptability in the model.

Comprehensibility and approachability. Another pattern in the dataset related to evaluations of comprehensibility. The raters in this study drew on a much wider range of speech features than just pronunciation, noting that grammar, vocabulary, and fluency features also impacted the ease of understanding the test takers in the samples. This finding is in line with previous findings on comprehensible speech (Crowther et al., 2022; Saito et al., 2016; Trofimovich & Isaacs, 2012). Also of interest is that the raters often explicitly mentioned that L2 accents did not play a large role in how easy or difficult they found speech to understand, which is also in line with past research (Munro & Derwing, 1995; Trofimovich & Isaacs, 2012). However, the raters in this dataset also drew extensively on nonverbal

behavior and affect when describing speech that was more or less comprehensible.

The raters broadly described a positive effect of approachability—broadly conceived as personability and presence—as enhancing comprehensible speech. The lack of approachability often inhibited the raters' ability to understand speech as easily. Approachability included elements that conveyed positive affect, such as smiling, laughing, and overall facial expressiveness. Positive affect helped create a sense of personability and rapport with the raters, which the raters reported as making them want to communicate with the test takers. These positive behaviors provided an important point of connection between the two individuals even despite the fact that the samples were recorded and not live. Engagement was also a critical element of approachability, as the constituent behaviors conveyed a sense of presence and desire to communicate on the part of the test taker. Being attentive by displaying mutual gaze to the examiner during test questions was one behavioral aspect of this, as well as using interactional behaviors such as nodding. In these cases, comprehensibility was enhanced even when the raters noticed weaker vocabulary, grammar, fluency, or pronunciation. On the other hand, inexpressiveness in both paralinguistics (e.g., monotone, flat prosody) and embodied behavior (e.g., rigidity), neutral affect, erratic gaze patterns (shifting, gaze darting around), and evidence of anxiety all contributed to lower comprehensibility in the test takers. The dataset suggested that an absence of movement overall resulted in raters feeling detached from such responses.

The finding that behaviors influenced comprehensibility aligns with recent past research in this area. Nagle et al. (2022), in a study of affect, found that greater collaborativeness (a measure of engagement) and lower anxiety were associated with higher comprehensibility scores. Tsunemoto et al. (2022) considered specific nonverbal behaviors and their impact on comprehensibility, finding that averted gaze associated with higher comprehensibility scores, which may have been due to the participants engaging in thinking rather than a lack of attention. They also found that greater eyebrow expressiveness corresponded with lower accentedness and higher fluency scores. Trofimovich et al. (2021) also found that nods, a nonverbal backchannel associated with engagement, led to higher comprehensibility scores, and positive affect played a minor role in enhancing comprehensibility in a subset of the raters. Finally, Kim et al. (2023)

studied the impact of behaviors on fluency ratings (which may be considered a constituent aspect of comprehensibility), finding that smiling and eyebrow movements led to greater fluency scores. As a whole, these elements of engagement and positive affect align with the elements of approachability that the raters in this dissertation study reported.

Overall proficiency and assuredness. The final pattern related to the relationship between assuredness and holistic impressions of language ability. Assuredness was mostly conceived of as confidence, but it was closely related to anxiety as well. The raters' decisions in this area were both based on an overall perception of affect and affect as perceived through nonverbal behavior. In particular, confidence was often perceived as a lack of behaviors that showed signs of struggle, such as shifting and averted gaze or being perceived as searching for linguistic resources. The presence of such behaviors led to direct judgements of low confidence. Inexpressiveness in facial expressions or otherwise non-target-like, "awkward" movements (e.g., fidgeting, self-adaptors) also contributed to these judgements. Averted gaze, if perceived as a sign of thinking about content and preparing an upcoming utterance, was not perceived negatively. Being perceived as confident and at ease almost always associated with positive impressions of language ability, often described as fluency in the broad sense of the term. Nonetheless, though evidence was limited, it did not appear that being anxious or lacking confidence necessarily *caused* lower proficiency scores; some raters mentioned that a lack of ability would lead to more anxiety and lower confidence in a testing situation. Thus, raters viewed proficiency and assuredness as being closely interrelated aspects of the same phenomenon.

However, given that many of the comments regarding assuredness also related to competence, and given that competence was also at least partially related to successful interactions, assuredness may have provided raters with additional evidence of overall successful performance outcomes. Supposing this is the case, these findings suggest that not only is confidence an integral predictor of educational achievement (Ahammer et al., 2019; Cobb-Clark, 2015; Heckman et al., 2006; Judge & Hurst, 2007; Stankov et al., 2012), but the *perception* of confidence also becomes part of the *perception* of competence. In relation to overall proficiency, the close relationship between assuredness and overall score impressions relates to

theorizations in the L2 literature that confidence and low anxiety associate closely with gains in language acquisition (Botes et al., 2020; Clément et al., 1980; Clément et al., 1994; Clément & Kruidenier, 1985; Dewaele & Alfawzan, 2018; Dewaele & Li, 2022; Dewaele & MacIntyre, 2014; Dewaele et al., 2019; Doqaruni, 2015; Jiang & Dewaele, 2019; Jin et al., 2017; Labrie & Clément, 1986; Li et al., 2020; MacIntyre et al., 1997; Noels & Clément, 1996; Noels et al., 1996; Teimouri et al., 2019).

Triangulated findings

The final stage of explanatory sequential mixed methods designs is to triangulate the findings from the various components (Creswell & Plano Clark, 2017). A triangulation analysis gives information about the overlap of key findings, with more overlap indicating a greater confidence in the findings. The three studies brought forward evidence that largely aligned together and coalesced into three general findings.

Comprehensibility, engagement, and positive affect

The first triangulated finding was that comprehensibility was measured not only by linguistic elements of language (fluency, vocabulary, and grammar), but also by affective stances that were determined by seeing nonverbal behavior. Study 1 showed that involvement was a key predictor in changes in comprehensibility scores. Involvement was made up of engagement, attention, and interactiveness, and as such can be seen as a measure of presence in the test interaction. Study 2 showed that mean valence predicted comprehensibility scores when base proficiency level was taken into account. Lower-level candidates were more likely to benefit from higher valence than candidates with higher proficiency. This showed that iMotions measurements of positive affect (e.g., smiling) corresponded to greater comprehensibility. Finally, study 3 demonstrated that approachability was a main factor that raters took into account when using evidence to make decisions about comprehensibility. Approachability was made up of elements of engagement, confidence, low anxiety, and positive affect, with constituent behaviors such as smiling, attentive gaze, and forward leaning posture playing important roles.

This study provides ample support for previous research in the area. These authors have argued that affect (Nagle et al., 2022; Trofimovich et al., 2021) as well as discrete behaviors (Tsunemoto et al., 2022) add to a greater ease of understanding L2 speakers. Behaviors showing engagement (showing attention,

leaning forward, smiling) were consistently elements that the raters in this dissertation used when rating and mentioned as making speech easier to follow. Although positive affect was not a key predictor of comprehensibility in the first study, it was in the second and third, showing that a friendly, personable demeanor was important in facilitating communication, especially for speakers with lower proficiency. Positive psychology would support these findings given that positive emotions have a broadening and building effect, showing that happy and warm individuals are interested and have a desire to get involved (Fredrickson, 2001, 2003; Seligman, 2011). This was not always fully supported in the literature (Chong & Aryadoust, 2023, Trofimovich et al., 2021), where positive affect had a somewhat inconclusive effect on language outcomes. Indeed, as noted by the extremely small R^2 value in Chapter 6, it may be the case that positive affect has an impact, but it is quite small. Nevertheless, the evidence suggested that both engagement and positive affect quite possibly allowed test takers to establish rapport with raters, even through remote recordings, which may have had an impact on raters as an affective contagion (Elfenbein, 2014; Hatfield et al., 1994), leading the raters to correspondingly show engagement or positive affect. Having "contracted" the test takers' positive affect, the raters may have then wanted to listen and communicate with the test takers. With greater attention being paid to the test takers, the resulting speech would likely have been understood more easily, resulting in a higher comprehensibility score.

A question that remains is whether the narrative from positive psychology translates across cultural boundaries, and whether the findings from these raters would be generalizable to other cultural contexts. All the raters in this study were American by birth, and thus shared similar (though certainly not the same) cultural backgrounds. Wierzbicka (1994) noted that for middle-class (and generally, white) Americans, interaction may be characterized by "great emphasis being placed on being liked and approved of, on being perceived as friendly and cheerful" (p. 182). Happiness may communicate success and achievement, pride, superiority, and self-esteem in American participants (Uchida & Kitayama, 2009; Kitayama, et al., 2006; Shaver, et al., 1987). Thus, test takers smiling, laughing, and otherwise being perceived as exhibiting positive affect may have unconsciously communicated to the American raters a certain degree of comfort and success during the speaking test. While features such as fluency, grammar, and vocabulary had features

that stood out and could be detected (relatively) easily, comprehensibility may have served as a catch-all category that indicated overall success in these circumstances, thus being more influenced by the approachability of the test takers. Whether these affective stances would have communicated the same type of information to raters from other backgrounds is thus unknown at this point.

Proficiency outcomes and confidence

Confidence also emerged as a strong predictor of proficiency outcomes—in particular fluency, vocabulary, and grammar—across at least two of the three studies. In study 1, assuredness, a measure of confidence and low anxiety, predicted changes in these three proficiency measures but not comprehensibility. In study 3, confidence was a frequent topic of the raters' thought processes and related quite closely to perceptions of language ability. Both of these studies suggest a bi-directional relationship for confidence (Edwards & Roger, 2015) in which greater confidence may make someone appear more proficient, and greater proficiency may make someone appear more confident. The raters used confidence as cognitive evidence of ability, as it related to the latent trait of proficiency, and also as a psychological affective measure, as it conveyed social information during the testing context. The various roles of confidence align with arguments made by Stankov et al. (2012), that confidence plays various cognitive, psychological, and social roles in interaction. In study 2, proficiency outcomes were positively predicted by variance in eye gaze in the group with higher proficiency, while greater variance in gaze associated with lower scores in the lower proficiency group. In the higher proficiency group, gaze may have appeared more target-like, shifting back and forth naturally similarly to proficient communication between L1 speakers (Goodwin, 1980; Rossano, 2012). More natural shifts in gaze may have conveyed attention and also a sense of confidence to the examiners, as a key function of gaze was to deduce confidence and anxiety in the rater group. Indeed, in Tsunemoto et al. (2022), the frequency counts of averted gaze associated with greater comprehensibility, which may have also been a measure of confidence in a similar vein. In the lower proficiency group, higher variance in gaze patterns may have been due to shifting gaze during moments of breakdown and struggle (such as in Burton, 2023). As noted in the recalls, shifting gaze was unanimously perceived as negative as it was seen as a sign of anxiety. It is possible that this finding from study 2

regarding attention shifts indirectly supported the findings about confidence here.

The close relationship between assuredness and proficiency outcomes is widely attested in the literature in both L2 studies and studies of general achievement (Ahammer et al., 2019; Botes et al., 2020; Clément & Kruidenier, 1985; Clément et al., 1980; Clément et al., 1994; Cobb-Clark, 2015; Dewaele & Alfawzan, 2018; Dewaele & Li, 2022; Dewaele & MacIntyre, 2014; Dewaele et al., 2019; Doqaruni, 2015; Heckman et al., 2006; Jiang & Dewaele, 2019; Jin et al., 2017; Judge & Hurst, 2007; Labrie & Clément, 1986; Li et al., 2020; MacIntyre et al., 1997; Noels & Clément, 1996; Noels et al., 1996; Stankov et al., 2012; Teimouri et al., 2019). This study thus adds to these findings in that the affective stance of showing confidence and low anxiety can impact how someone is perceived. This finding may also help explain the findings in Jenkins and Parra (2003) and Neu (1990), case studies in which more confident test takers were able to overcome weaker linguistic skills when displaying a greater amount of assuredness.

Listening comprehension and competence

There was also some convergence in the importance of listening comprehension and how it factored into scores. Study 3 found that raters paid close attention to listening comprehension, frequently remarking on the presence of nonunderstanding and repair. They then observed whether test takers were able to repair and reestablish intersubjectivity or mutual understanding (Burch & Kley, 2020). Being able to realize successful interactions was thus an important criterion for the raters. Raters often discussed comprehension, repair, and success in terms of competence, one of the scales used in the rating study. Interestingly, study 1 reported that competence correlated with all four proficiency outcomes (M = .76), with an association strong enough to support its inclusion in a language factor along with the four proficiency outcomes. This shows that the raters implicitly viewed listening comprehension as a core aspect of spoken communicative competence, even though it was not included explicitly in the rating scales. A core role for input comprehension is also supported by literature on psycholinguistic models of communicative competence (e.g., de Jong, 2023) and work on speech processing (Levelt, 1993), which may relate quite closely to semantic prediction (Levinson, 2016; Pickering & Garrod, 2013).

In studies on the relationship between different constituent elements of fluency (speed, breakdown,

and repair; Tavakoli & Skehan, 2005), it is notable that repair fluency rarely correlates highly with fluency measures when taking into account speed and pausing (Saito et al., 2018; Suzuki & Kormos, 2020). If raters truly consider listening comprehension as integral to L2 fluency, one would expect that repair fluency would form a stronger relationship with fluency measures. One possibility to explain this attenuated relationship may be in how test takers manage breakdown sequences. Findings from study 3 suggested that more adaptable test takers may leverage their affect and nonverbal behavior to manage repair sequences more effectively than less adaptable candidates. Adaptability would then moderate the relative impact of comprehension breakdowns and repair and could then explain why repair fluency may play less of a role in how fluency is perceived.

Overall impact of nonverbal behavior

The three studies showed that nonverbal behavior, either viewed through its semantic, cognitive, affective, or social roles or as a discrete phenomenon, had an impact on all four proficiency outcomes in different ways. What is of interest in the overall size of the impact. In study 1, relationship between subjective, observed predictors and outcomes explained about 20% of the variance in scores. In study 2, the relationships between externally measured, objective predictors were much smaller, explaining around 2% of the variance in scores. In study 3, nonverbal behaviors made up 11% of the total number of comments. The literature shows similar types of relationships. For comments in stimulated recall designs, Sato and McNamara (2019) and May (2011) found a similar number of comments about nonverbal behavior in relation to communicative success and interactional competence, respectively. Choi (2022), Nakatsuhara (2021a), and Nambiar and Goon (1993) all showed that the visual realm exerted a positive impact on test scores, approximately enough to raise scores by half a band in the context of IELTS. Trofimovich et al. (2021) reported that adding in a predictor of nodding resulted in an explanatory gain of about 13% in their model of comprehensibility. Chong and Aryadoust (2023) did not report variance explained, but instead noted that it was too negligible to support claims that the models were meaningful. Tsunemoto et al. (2022) did not report R^2 values or standardized effect sizes in their modeling of behavior, unfortunately. In any case, although these various values are not directly comparable, they appear to converge in that nonverbal

behavior is able to explain variance in each of these studies, with score gains being noted by all authors. However, the size of those gains is not always enough to change scores substantially. Regardless, in a high-stakes situation, *any* variance explained, even if it is 3%, may be enough to shift outcomes for particular test takers. This alone is reason to investigate the phenomenon, consider construct revision, and make efforts in rater training. I now turn to issues with the construct in the next section.

A revised model of L2 communication

This findings from this dissertation suggest that our understanding of communicative competence, which informs models of language proficiency, may only partially explain how L2 speakers use language in real world settings. A tiered visualization of these elements was presented in Figure 2.1, organized in layers ranging from those located within the individual at the center and those being co-constructed with others in the outer layers. At the core of this model are cognitive elements, which represent the inner workings of language comprehension, processing, prediction, and formulation. Structural elements are the traditional linguistic competences of grammar, vocabulary, and pronunciation. Discursive elements are those cohesive and organizational features that help utterances create complex meaning. Structural and discursive features are learned, and cognitive features represent the automatization of their acquisition. Co-constructed elements are those that the user creates with others depending on situational needs, including strategic, pragmatic, and interactional elements. These are leveraged according to context and allow the speaker to make meaning with others. These competences are situated in and react to social context in order to fulfil communicative aims. Missing, however, are the effects of nonverbal behavior and affect. If positive affect, engagement, and attention can indeed play major roles in meaning making with others, and if constituent nonverbal behavior reveals something about the cognitive, structural, and organizational features of language, a revision is necessary to better encompass these elements. These elements are critical for ability for use, a core aspect of Hymes' (1972) communicative competence, which included these psychosocial aspects of communication. Any model eschewing the visual realm would be logocentric (Mondada, 2016), failing to represent the true complexity of human communication.

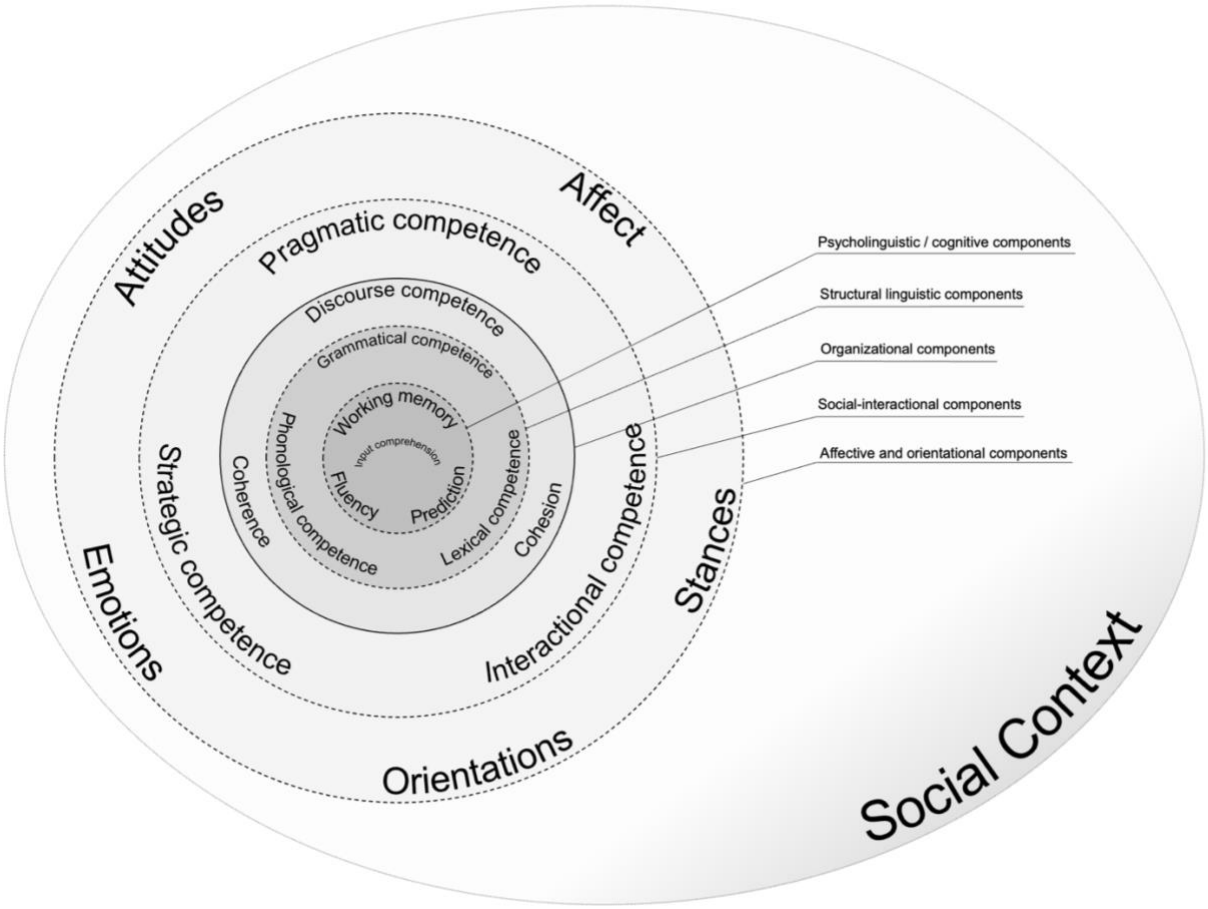
As reviewed in the literature, nonverbal behavior was described as playing a primarily social-

interactional role in prior models of L2 communicative competence. Canale and Swain (1980) and Canale (1983) described how nonverbal behavior played a compensatory role in strategic competence, and Celce-Murcia (2004) and Galaczi and Taylor (2018) more explicitly included a range of nonverbal behavior in their discussions of interactional competence. However, I reviewed studies from psychology, applied linguistics, and human communication showing that nonverbal behavior can also convey cognitive, semantic, social-interactional, and affective information. L2 communication can combine verbal and nonverbal elements to provide information about the language development of the speaker and their ideas and intentions. Nonverbal and verbal strategies and interactional moves help speakers navigate breakdowns, compensate for gaps in their lexical knowledge, and manage conversation with others. Conveying affective information allows speakers to demonstrate their inner stances, motivations, desires, and feelings about events, which can provide important cues for interactants navigating dynamic and fluid contexts. The deployment of pragmatic competence, interactional competence, strategic competence, and skillful affect is the hallmark of highly competent L2 speakers (Roever, 2021), and when raters view this, they may perceive an individual as stronger than their actual inner psycholinguistic structural and cognitive competences. Affect and emotion are furthermore co-constructed amongst interlocutors, and the affective behavior of one person can impact the responses of another in different ways. These responses can also color the judgements people make about their interlocutors.

In the studies discussed in this dissertation, I have presented additional evidence that affect and nonverbal behavior can further color perceptions of the language abilities of speakers. Perceived assuredness provides clues as to speakers' underlying proficiency (cognitive traits, such as fluency, grammar, and vocabulary). Being perceived as engaged (through nodding and leaning forward, for example) may engage others in closer listening behaviors, thus improving their comprehensibility, perhaps even compensating for developing phonemic control. Eye gaze patterns marked by a greater number of shifts may indicate struggle in lower proficient speakers and provide evidence of lower cognitive fluency, while positive behaviors such as smiling aid in comprehensibility. A full range of other behaviors that raters observed provided evidence for other traits relating to language proficiency as well. For this reason, there

is strong evidence to revise models of L2 communication to incorporate nonverbal behavior, affect, and the ever-dynamic role of context. Figure 8.2 displays an extension of de Jong's (2023) conceptualization of language proficiency, presented earlier in Figure 2.1, adding in an affective dimension that closely interacts between language abilities and social context.

Figure 8.2
An Extended Model of Language Proficiency

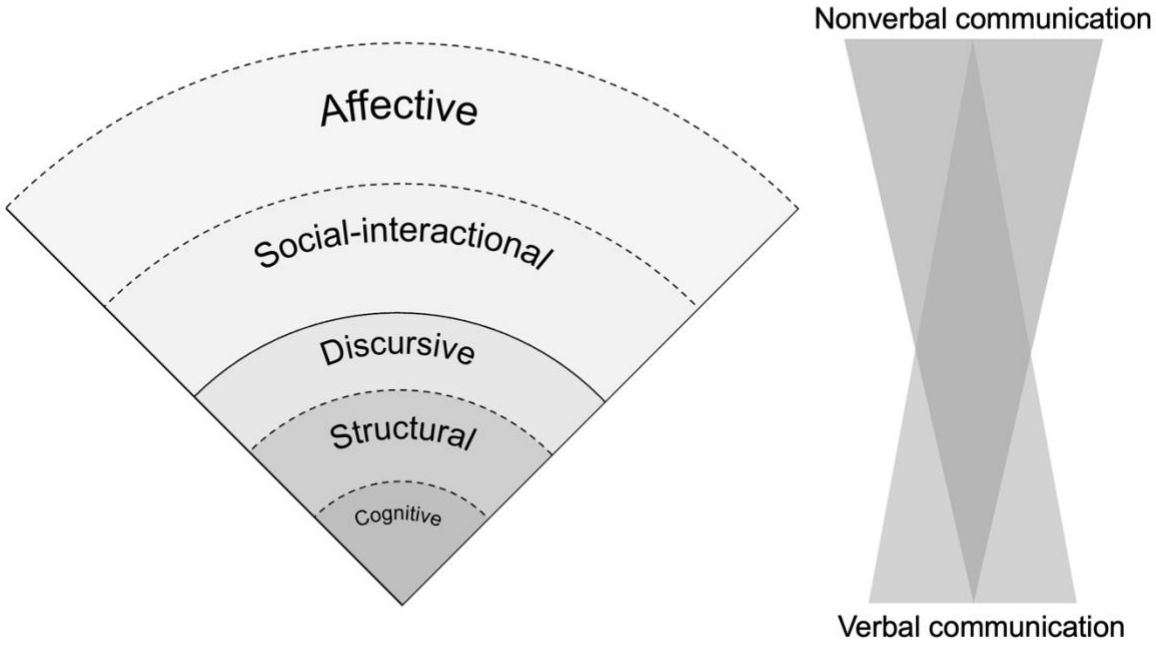


The abilities of individuals to convey affective and emotional meaning in the outside tier of this model are culturally bound and largely innate. As such, I do not classify them as competences. Rather, together with strategic, interactional, and pragmatic competences, they are socially co-constructed aspects of meaning making that contribute to overall messages, may reveal information about a person's language ability, and add to their capacities to communicate. These external layers allow users to adapt to various social contexts with varying task demands, goals, and purposes. They can mediate language when there are

breakdowns in understanding or ability, allowing speakers to succeed even in the face of challenges. Alterations in the valence and strength of affect have been shown to moderate listeners' interpretations of speech, and as such, their inclusion in the model is warranted.

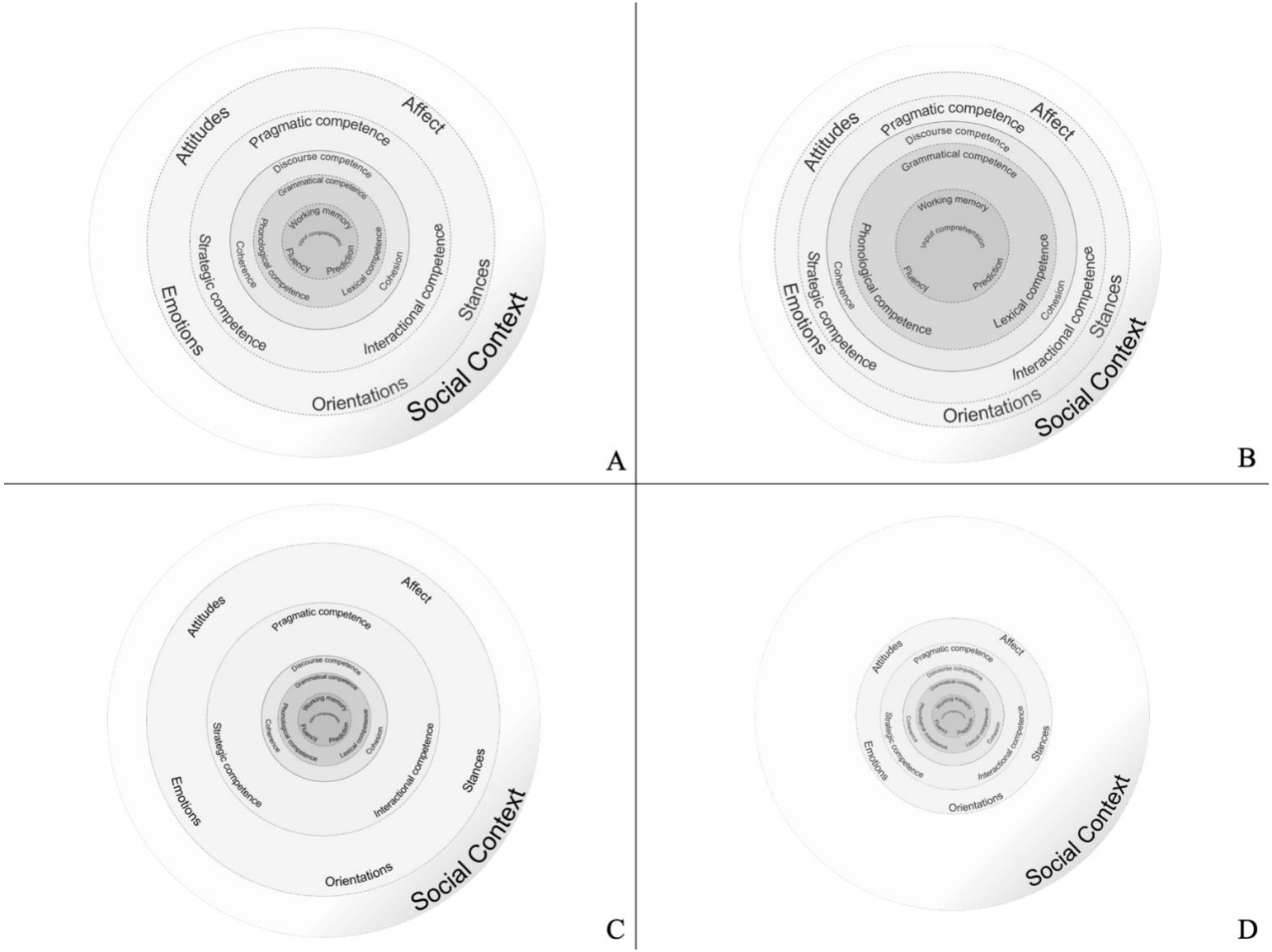
My interpretation of the relationship between nonverbal and verbal communication as it relates to the model above is presented in Figure 8.3. Verbal communication generally conveys semantic, ideational meaning, and it is a core component of cognitive and structural aspects of speech. It can, however, also convey meaning about people's affective responses, stances, and orientations. Nonverbal communication is generally non-propositional, and thus the information it conveys is largely not semantic in nature. Instead, nonverbal behavior conveys large amounts of socially oriented information. It conveys affect and orientations to others, as well as aiding in social situations by managing interactions and providing strategic resources for speakers. That being said, nonverbal behavior may be used to infer information about speech fluency and comprehension (cognitive components), as well as lexicosyntactic and phonological information (structural components). The two modes of communication are linked, and meaning is often combined from both modes to convey meaning from all levels. This is conveyed in Figure 8.3 by the overlapping triangles, which show the strength of the associations with each aspect of proficiency.

Figure 8.3
Relationship Between Verbal and Nonverbal Communication



Past models and theorizations of communicative competence have often described target-like "effective" communication, but these have not made a strong case for including aspects of language development in their theorizations. Current models describe competence within the confines of an idealized, often L1 "native" speaker. These can be problematic as L1-like attainment can be unrealistic for learners, and an undesired goal, as bi- or multilingualism can be an important part of one's identity (De Costa, 2016; Norton, 2000, 2013; Pavlenko & Norton, 2007). Learners may be competent in a second language and communicate effectively at many different levels of development. What is necessary, then, is a model which is flexible enough to allow descriptions of the cognitive, structural, discursive, and social-interactional patterns of development that have been discussed in second language acquisition research. The visualization presented above, with the integration of psycholinguistic core aspects of proficiency with social-interactional and affective external aspects, provides a useful metaphor for understanding how profiles of language learners with differing learning trajectories interact with social context. These are presented in Figure 8.4, and described below.

Figure 8.4
Profiles of Learners with Different Competences and Skills

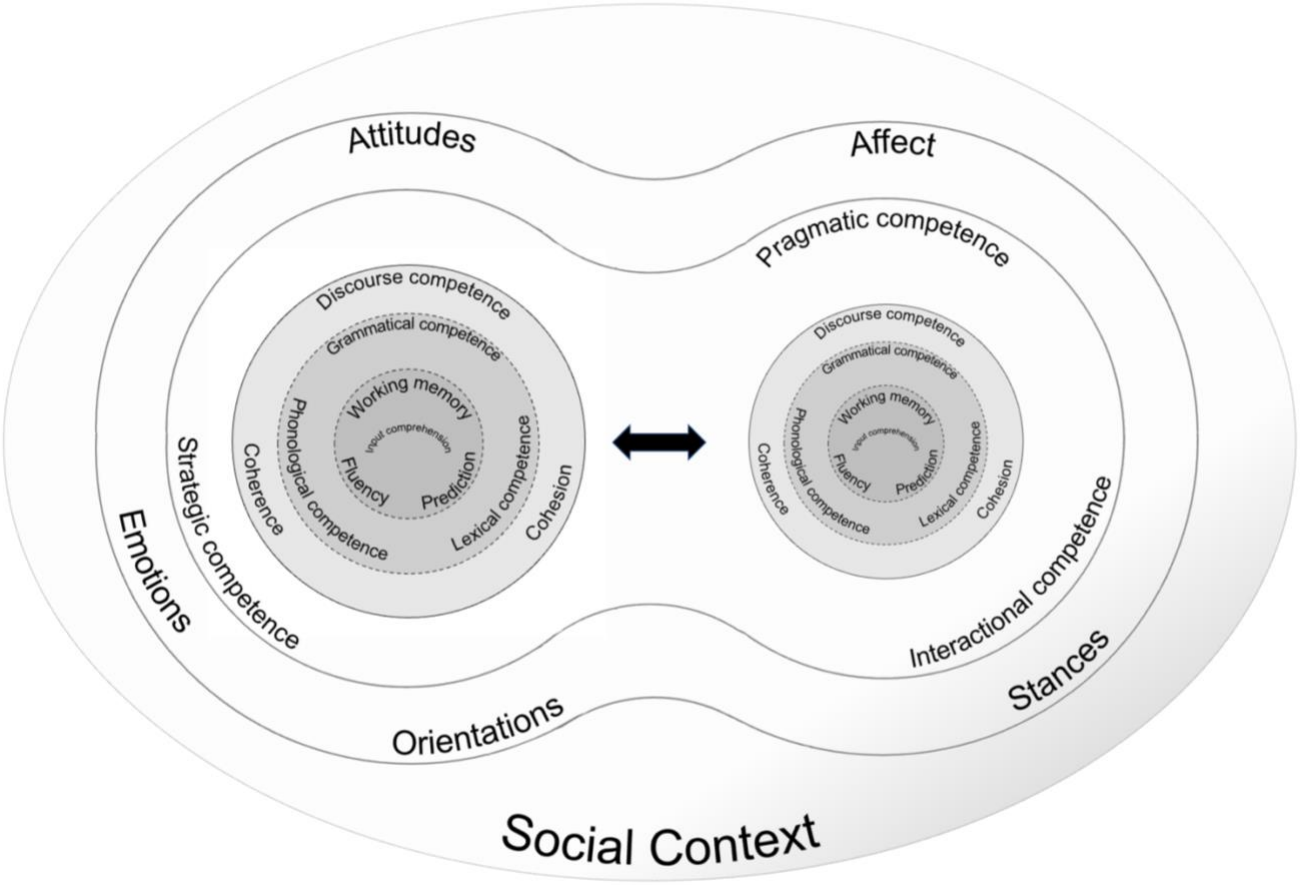


In Figure 8.4, hypothetical person A demonstrates an L2 learner with a balanced profile of the various competences and skills. Each tier is approximately the same size, and the ability overall covers nearly all of the needs of the social context. This speaker would be able to fulfill the requirements of social interaction relying on a mix of all of their competences and skills, thus being perceived as an effective speaker. Person B is also an effective speaker and can fulfill the needs required by social context. This profile, however, shows a stronger core of psycholinguistic aspects of language learning. This person is able to rely on their knowledge of grammar, vocabulary, and phonological features, because they are automatized and available for use. In this particular social context, the individual does not need to leverage interactional or affective aspects of communication as much to meet the demands of context. Person C, however, is quite different. They are also able to successfully communicate, but they do so with a much more reduced set of inner linguistic competences for the context. Instead, they are able to leverage their pragmalinguistic tools, strategic competence, interactional abilities, and affective stances to communicate effectively despite their relatively weaker linguistic skills. The differences between individual B and C may represent why differing profiles of test takers in Jenkins and Parra (2003) and Neu (1990) were perceived differently in their test discourse. Finally, person D has a much more reduced set of competences and skills, and these are not sufficient to meet the needs of social context. The individual is not able to leverage interactional and affective stances in the same way as person C.

In my theorization in this model, none of these individuals have a static set of competences and skills. These change and morph according to the needs of social context. Thus, these can also be thought of the same individual in different contexts, as each situation will have a different set of needs that requires experience and learning to accomplish. Not every situation will accommodate the affordances of affective stances that are primarily visual, such as telephone conversations. Likewise, situations with different levels of difficulty or power dynamics will reshape and modify the competences and skills a speaker will be able to display.

This type of metaphor of language ability may also be able to extend to interactions themselves. A foundational argument regarding interaction is that speakers co-construct discourse (Galaczi & Taylor, 2018; Roever & Dai, 2021; Roever & Kasper, 2018; Plough et al., 2018; Young, 2011), with each speaker exerting an influence over the other. Disentangling the language abilities of more than one speaker, especially interactional competence, can therefore prove complex. The above metaphors may offer a useful way to visualize these effects as well. Figure 8.5 shows how interactional-facing skills may appear in a dyadic encounter. Here the person on the left has a somewhat larger language profile in terms of core elements for the particular social context, while the person on the right has a somewhat smaller profile. These are separate because they reside within each individual. The social-interactional and affective elements of these two individuals, on the other hand, are shared. The dyad co-constructs their encounter, making meaning using their interactional competence and impacting each other through their affective stances and orientations. In this particular example, the interactional and affective domains are approximately equal in size, and together the two individuals cover the needs of the social context. They are thus able to communicate successfully in this social encounter. In other encounters, the interactional and affective contributions from both speakers may be somewhat different in size, but they will still overlap. They may not cover all of the contextual requirements of the situation, and communication may not be as effective. This metaphor could also be useful for understanding sociocognitive theories of language learning (Lantolf et al., 2020; Vygotsky, 1978). In these theories, a stronger interactant may scaffold language, affect, and elements of the social context to enable less proficient learners to achieve communicative goals that are normally beyond what they can do alone or with a weaker interactant. The stronger interactant can, as one might say, “bring out the best” in the test taker.

Figure 8.5
Language Ability in a Dyadic Encounter



The importance of the socially oriented layers of competence in the discussed models has been attested especially with regard to advanced learners. Roever (2021) made the case earlier that pragmatic knowledge is especially important for learners above at and above B2 level, as the nuances of communication become much more than conveying the meaning of particular words and grammatical structures. Importantly, Roever (2021) also noted that nonverbal affective information can help compensate for any gaps in a speaker's pragmatic knowledge, helping to convey meaning the verbal channel cannot. Lantolf (2006) made a similar argument regarding the unification of language and culture into a broadly termed "languaculture", which becomes increasingly important for learners to tap into as they grow in language proficiency. He used a similar circle metaphor, referring to the outer circle in terms of cultural elements corresponding with language:

> Outside of the circle, the domain of languaculture, meaning becomes much more interesting and complex because it entails knowledge of different concepts and how these are encoded in such features as conceptual metaphors, lexical networks, lexicogrammatical structures, schemas and the like that represent different ways of organizing the world and our experiences in it. (Lantolf, 2006, p. 79).

Essentially, Lantolf is describing similar elements in the models above, referencing pragmatic competences and interaction. He goes on to emphasize nonverbal behavior as well, referencing Slobin (1996, 2006) and Talmy's (2000) work on path gestures to highlight aspects of culture visible through path motion depictions of manner are critical to conveying meaning, yet lie beyond the inner psycholinguistic core of language proficiency.

Interaction is likely the primary driver of perceived advanced language proficiency. The fundamental importance of interaction is well attested in the literature in SLA (Gass & Mackey, 2020; Long, 1996), which drives acquisitional processes when individuals need to use language in a social context. Interaction is fundamental as well to sociocultural theories of language acquisition (Lantolf et al., 2020; Vygotsky, 1978), as interactants can scaffold the social context creating opportunities for learners to use language beyond what they would normally be able to produce without support. Interaction creates opportunities for speakers to demonstrate the widest range of their abilities, include pragmatic, strategic, and interactional skills (Roever, 2021). Interaction, when built into language assessment, can also create an environment where learners feel more at ease and demonstrate a wider range of demonstrable language (Thompson et al., 2016). As nonverbal behavior is critical to interaction management and conveying culture (Celce-Murcia, 2007; Galaczi & Taylor, 2018; Lantolf, 2006; Plough et al., 2018), narrowing language performance to audio-only or non-interactive formats would limit the abilities of test takers to display their full range of language proficiency. This would further explain score differences found in Choi (2022), Nakatsuhara et al. (2021a), and Nambiar and Goon (1993) where the visual realm was stripped from audio-only rating, and test takers were not able to show their full L2 proficiency.

Implications

Implications for language testing and SLA research

Researchers have argued for decades that nonverbal behavior and affect may exert an effect on performances or scores in speaking test settings (Pennycook, 1985; Harding, 2014; Kellerman, 1992; Plough et al., 2018; Plough, 2021; Young, 2002). These non-linguistic elements have regularly been the "elephant in the room," as practitioners witness these effects in operational settings, but the score impact has largely been unknown or only posited through small scale studies. This dissertation has then provided some initial evidence of impact across a larger sample of individuals in a language testing context. It has shown that novice raters that are not trained on specific rating scales take nonverbal behavior and the information it conveys into account when formulating impressions of communicative competence, and these become part of score variance, albeit to a somewhat small degree. It thus confirms many of the findings of Gan and Davison (2011), Jenkins and Parra (2003), and Neu (1990), that nonverbal behaviors can have an impact on score outcomes. It also explains some of the variance in test scores due to modality differences (e.g., Choi, 2022; Nakatsuhara et al., 2021a; Nambiar & Goon, 1993).

Theory to date has built a solid argument that affect and nonverbal behavior are integral parts of human communication (Hall et al., 2019; Hall & Knapp, 2013; Matsumoto et al., 2016). It has even been speculated to belong to an expanded version of communicative competence (e.g., Canale & Swain, 1980; Hymes, 1972). Test constructs that draw from logocentric theorizations of L2 communication (Mondada, 2016)—those that do not include nonverbal behavior—may suffer from construct underrepresentation (Messick, 1989) by not including critical facets of the test construct that exist in the target language use domain. Indeed, apart from few accounts (such as the test revision project described in Jenkins & Parra, 2003), large-scale tests (like IELTS or TOEFL) rarely include descriptors of nonverbal behavior beyond paralinguistic cues (e.g., prosody and pauses). In some testing contexts, developers have chosen audio-only recordings as the basis for rating speech, thus entirely removing the visual world from the speaking test. These performances often result in scores that are lower than performances in which the test taker can also be seen (Choi, 2022; Nakatsuhara, 2021a; Nambiar & Goon, 1993). Removing the visual world can be

problematic if it disadvantages test takers by not accounting for their repertoires of communication that go beyond those that are verbal only. One way to strengthen the construct would be to include video recordings of test takers as the basis of rating (or live, face-to-face rating), and to include a broader range of criteria that reflect this expanded construct. As Plough (2021) argued in regard to the inclusion of nonverbal behavior in test constructs, "we are obligated to create rubrics that, to the extent possible, account for the full range of performance (on which candidates are evaluated)" (p. 62). She went on to issue a word of caution given the challenges of individual variation, idiosyncrasy in interpretations, and contextual fluidity of nonverbal behavior and the meaning it conveys. More research is needed in this area to determine which aspects of behavior give reliable insight into language ability across cultures and contexts, and whether these are meaningful in terms of second language development.

The results of this study have implications for discrete skills and integrated skills assessment as well. In the tradition of discrete testing, reading, writing, listening, and speaking are tested separately, and efforts are made to minimize the impact each has on the other. In productive skills testing, however, input is necessary to elicit language from test takers (e.g., a prompt, a question from an interlocutor), and decoding that input always requires receptive skills of listening or reading. In the case of speaking, input is generally in the form of some aural stimulus, such as a conversation partner, though it can also be in the form of written instructions, or mixed (multimodal). This format of testing requires speakers to both listen and speak in productive skills tests, but they are only scored on the basis of their spoken performance. The paradigm of integrated testing, on the other hand, treats skills as interrelated and inseparable. Listening and speaking form part of a unified construct, and scores provide inferences about both skills. The results from this study suggest that raters naturally find listening to be a core part of the speaking construct, aligning with past findings (Brown et al., 2005; Jenkins & Parra, 2003; May, 2011; Orr, 2002; Sato & McNamara, 2019). It may be necessary to represent listening comprehension in rating scales to avoid construct underrepresentation, thus more broadly adopting a form of integrated assessment. Given the wide range of verbal and nonverbal features the raters mentioned, and the prevalence of these in the literature, it may be possible to begin devising more meaningful scale categories for listening comprehension in speaking tests.

Other aspects of nonverbal behavior and affect may merit their inclusion in other scales, in particular fluency. For example, the presence of shifting gaze, self-adaptors, and relative inexpressiveness corresponded frequently with speakers with more limited language skills. A more attentive gaze (not purely defined by mutual or averted gaze), the use of co-speech, representational gestures, and a more skillful use of head nodding and eyebrow movements corresponded with more proficient speakers, as well as speakers that conveyed a greater level of comfort and confidence. I do not think these behaviors should be assessed discretely or separately from language subskills (see Jungheim [2001] and Pan [2016] for examples of the discrete assessment of nonverbal behavior that are particularly problematic). As O'Sullivan (1996) argued after discussing his own efforts to devise such an assessment of nonverbal competence for L2 speakers,

> [t]hough the possibility of developing tests which will indirectly test such competence is certainly appealing, it is as inappropriate to separate the non-verbal channel from its natural context of communication as it is to separate the verbal channel. Therefore, in as much as previous tests can be argued to lack validity for ignoring one important aspect of communication, such indirect tests will lack validity for the same reason (p. 319).

The results in this dissertation align with the above comment in that raters take a holistic view of nonverbal behavior and integrate evidence from verbalizations when making their decisions. Thus, scales that include both verbal and nonverbal behavior may be more informative and useful for raters, thus providing a broader source of evidence about skill development. Fluency scales, for example, could include information about gaze and gesticulation. Vocabulary or grammar scales that include descriptors pertaining to the use of representational gestures may also be useful for raters when scoring these areas. Comprehensibility scales could include descriptors about behaviors that convey engagement and positive affect. If used, the wording of these should emphasize the skillful use of these behaviors when conveying meaning effectively *together* with language elements. Any use of descriptors describing the mere presence or absence of behaviors would undoubtedly poorly represent the construct, as raters do not consider behavior in a binary, yes/no manner.

Another important takeaway from this study is the need to strengthen rater training programs with discussions of behavior and affect. Unfortunately, the content of these trainings is rarely discussed in the

literature beyond methodological aspects (e.g., frequency/duration of sessions) (e.g., Yan & Chuang, 2023; Weigle, 1994, 1998). It is unknown how large scale or local language testing organizations address behavior and affect in the context of rating speaking tests. In my own experience having worked as a rater for multiple large scale testing organizations, I can attest that this type of training was extremely limited or non-existent. What may be needed then is ethics and sensitivity training. These sessions could focus on building empathy with test takers, as the testing situation can induce anxiety in many individuals and perhaps change the way that these test takers appear visually. These trainings would be especially useful for dealing with test takers with varying neurodiverse profiles who also have physical behaviors as symptoms of those profiles, such as rocking, repetitive movements, or differing patterns of gaze (American Psychiatric Association, 2013). It would be critical for raters to understand that individuals may exhibit differing patterns of openness and attention if they are, for example, autistic, and that these behaviors should not impact the rating of their language. This would require testing organizations to recognize the needs of these underserved groups to build equity and reduce bias (Randez & Cornell, in press). Nonetheless, the accommodations needed for many groups of test takers is still an active area of inquiry in the field (Taylor & Banerjee, in press), and much more research is needed to uncover how best to serve these varying groups.

In terms of SLA research, this study provides some evidence that there may be developmental trends in how speakers use gaze and gestural behaviors as they develop in particular areas such as fluency and comprehensibility. As discussed previously, it has provided some confirmatory evidence for studies in this realm such as Kim et al. (2023), McDonough (2019, 2023), Nagle et al. (2022), Trovimovich et al. (2021), and Tsunemoto et al. (2022). Nonetheless, much more research is needed to understand whether these behaviors and the cognitive, social, and affective information they convey can be reliably separated into stages of growth. Given the greater amount of research targeting gestures in SLA and the somewhat contested findings about patterns of gestural development (Aziz & Nicoladis, 2018; Benazzo & Morgenstern, 2014; Gullberg, 1998, 2006, 2012; Krauss & Hadar, 1999; Laurant & Nicoladis, 2015; Nicoladis, 2007; Nicoladis et al., 1999, 2007; Sherman & Nicoladis, 2004), it may very well be the case that hard coded or linear patterns of development do not exist, much as is the case with many aspects of

language itself. In this case, understanding the various ways behaviors and behavioral ensembles are used in context to display effective communication skills may be the best route to describing patterns of development.

Also in terms of development, tracking growth in high level speakers (upper intermediate/advanced, B2–C2 on the CEFR) may not be entirely appropriate using traditional methods in SLA. Currently, language development in SLA studies is often tracked using complexity, accuracy, and fluency (CAF) measures (Ellis, 2003; Ellis & Barkhuizen, 2005; Skehan, 1998). These measures provide discrete snapshots of development during educational interventions so that researchers can see gains in one or all three of the measures, thus justifying certain methods of language learning. These have been used extensively in the field, and they have provided vast insight for understanding growth in lower proficiency learners. Tracking growth using these measures, however, does not take into account critical aspects of functional adequacy, or whether communication is effective and sufficient to complete certain tasks (Pallotti, 2021). The raters in this dissertation were especially attuned to communicative adequacy, which they often scored as *competence*. Taking into account features of language that lead to functional adequacy, such as pragmatic competence and interactional competence, may be especially important for more advanced speakers (Roever, 2021). This dissertation has also shown that effective use of affect management (adaptability, approachability) in the face of unpredictability can lead to differences in how learners may be perceived, with more capable learners leveraging multimodal resources to accomplish tasks. If CAF measures fail to show growth in higher proficiency learners, it may not be due to the intervention but rather the measures themselves.

Methodological implications

This study has a number of methodological implications as well. Firstly, to my knowledge, it is the first of its kind to use iMotions facial analysis software to extract measures of nonverbal behavior in a study of L2 communication. Although Chong and Aryadoust (2022) used FaceReader, a very similar application that extracts base emotions, their study did not focus on nonverbal behaviors but rather how emotional transfer may have impacted test scores.The employment of iMotions in this dissertation was useful because

its measures broadly aligned with the findings from the rating scales and the stimulated recalls, showing that cutting edge facial recognition technology may be used in meaningful ways in L2 research. Although expensive, the benefit of using this technology is that it dramatically reduces the workload necessary to measure nonverbal behavior. I did not report statistical information from the human-annotated ELAN transcripts in this dissertation, but as a contrast, the research assistants took many months to transcribe 30 two-minute speech samples fully, while iMotions took less than an hour. The speed of these tools is paramount for the study of larger samples of data.

This comes with an important word of caution. As noted in Chapter 6, the correlations between the iMotions variables and the observed variables were medium to low, even in the case of the very similar measures of positivity and valence. One of the issues in using these algorithms is that developers do not always fully disclose a) how the technology works, b) the specific facial movements it measures in its emotional indices, or c) how accurate the classification system is. It is unknown whether the software can accurately detect facial movements on the faces of individuals from varying cultural and ethnic backgrounds; that is, the demographic information about individuals the software was trained on is not disclosed. If classification accuracy is low, this will mathematically attenuate any correlations with outcome measures, possibly resulting in skewed results or type II error. Thus, the use of these systems in L2 research will require more validation work to understand the underlying features being measured and the systems' accuracy in doing so.

On a related topic, one theme that has arisen from this dissertation was the contrast between perceived and observed measures. Perceived measures in this study were the scores awarded by the undergraduate student raters on the affect rating scales, while observed measures were those that were annotated by iMotions and the human-based annotations reported in Chapter 7. Gullberg (1998) found clear differences between perceived and observed gestures and how these related to score differences. In her study, raters' perceptions of gestural use varied from manual annotations of gestures. Observation did not always align with perception. Regarding the impact on scores, apart from one category of annotated iconic gestures, only perceived gestures impacted scores that her raters awarded. In this dissertation, Study 1

showed that the perceived measures of affect predicted with a certain degree of strength changes in the four proficiency outcomes. Namely, assuredness predicted changes in fluency, vocabulary, and grammar, while involvement predicted vocabulary and comprehensibility. Study 2, on the other hand, showed that objectively observed behavior through machine learning could also be used to predict outcomes. In this case, variance in attention predicted changes in fluency, vocabulary, and grammar, and overall valence predicted comprehensibility. However, the models with observed variables explained far less variance (2-3%) than the perceived measures (15-25%). This aligned with the stronger impact of perceived affect measured in Nagle et al. (2022) and the weaker impact of observed features in Trofimovich et al., (2021), Tsunemoto et al. (2022), and Chong and Aryadoust (2022). To some degree, it appears logical that perceived features would relate more strongly to language outcomes: Raters observe both, and there is probably some degree of overlap in what they perceive. As far as I can tell, in all studies that used perceived variables, these were all measured simultaneously in the same rating session with language. A methodological implication here is that studies need to be carried out where affect or behavior and language proficiency are measured at different yet counterbalanced moments in time. Only by designing a study in this way can the relative impact be teased apart.

Finally, I also believe that this study has shown the value and strength of using emic, rater- and test taker-focused methods when studying nonverbal behavior. Using an ethnomethodological approach, including stimulated recall and multimodal conversation analysis, "explores the ways social actions are built by the participants, contingent on and indexical for the specifics in any situation" (Kasper & Wagner, 2018, p. 82). Plough (2021) advocated for these approaches in research and test validation studies given that nonverbal behavior "is not a static behavior that can be categorized; rather, it is part of a dynamic interactional process" (p. 62). Multiple examples of these dynamic processes were seen in Chapter 7 on nonverbal behavior and rater cognition. Currently, modeling techniques in statistics may be incapable of modeling such dynamic behavior that shifts according to context. While there is some promise in the use of machine learning and AI to study complex, dynamic phenomena, these require massive datasets that are impractical to build for such quasi-exploratory research. Thus, while the empirical stance I took in Chapter

5 (on affect and language proficiency) and Chapter 6 (on nonverbal behavior and language proficiency) was justified and has indeed shown patterns useful for the study of this phenomenon, these findings would be far less meaningful without the insights from the raters and also the transcripts of the speech samples presented in Chapter 7. Given the nature of nonverbal behavior to co-occur with language and convey cognitive, affective, and social information, the mixed-methods design using an ethnomethodological approach in this dissertation is especially appropriate for studying this phenomenon of L2 use (Hulstijn et al., 2014).

CHAPTER 9: CONCLUSION

This study has presented a comprehensive analysis of the impact of nonverbal behavior and interpersonal affect on L2 proficiency outcomes. It used a three-tiered design using mixed methods to triangulate findings, thus leading to more stable inferences. The study has broadly found that nonverbal behaviors and affect can impact proficiency outcomes in different ways. Desirable, communication-forward behaviors such as mutual gaze, nodding, leaning forward posture, and representational gestures can convey confidence, engagement, and positive affect, which lead to differential outcomes in ratings of fluency, vocabulary, grammar, and comprehensibility. The variance explained by these phenomena was rather small, as one would expect. However, even a small amount of variance could prove important when a test taker is near a meaningful cut-point on a high stakes test. Thus, I have concluded that the results of this study and the broader literature on nonverbal behavior and affect point to a need to revise models of communicative competence, and I have presented one such alternative model. I have argued that the results here could also be operationalized in speaking test constructs, with scores providing a much more valid inference about language ability.

Limitations

There are a number of limitations in this study. For one, any results must be interpreted within the context of the participants: young, L1 English speakers in America observing the speech and behavior of L2 English speakers from China. These effects may not be universal or generalizable to other cultural contexts. For example, global contexts that do not place such cultural capital in appearing happy, positive, and confidence—such as the case of the United States—may not show the same correlations with proficiency judgements or improvements in comprehensibility. Without studies that extend this research to other groups, it is unknown whether these effects may be generalizable more broadly. Likewise, it is difficult to disentangle the effect of nationality of the sample participants from the study, as all test takers were from the same cultural group. More research is needed to understand whether the background of the L2 speakers influenced perceptions of their nonverbal behavior.

A second limitation concerns the speech samples. While the pool of raters was sufficiently large to detect a number of rather small effects, the sample of 30 test takers was rather small. Although generally representative in terms of spread of ability levels, this sample lacked extremely weak or strong speakers. The set of samples also may have lacked variance in expressiveness in nonverbal behavior. Nonetheless, the sample size was both a result of what the test developer could provide, as well as a result of the power analysis. In the future, larger samples with broader ranges of behavior can be used to observe the impact of these behaviors on language ratings. Instead of drawing from testing contexts, which may attenuate the expression of strong emotions, it may be more desirable to have learners produce authentic, real-world recorded language that can then be rated. This would also enhance the ecological validity of the methodological design, relating more strongly to target language use in the real world.

Although the test takers had a shared cultural background, the samples varied in both interlocutors and topics. It is known that the verbal and nonverbal behavior of raters can impact test takers' performances (Briegel-Jones, 2014; Brown, 2003; Plough & Bogart, 2008), so it is unknown to what extent the raters in this study impacted the test takers' multimodal discourse. Likewise, the tasks varied in these samples, with some topics appearing somewhat more difficult than others. Breakdowns in comprehension did not appear in all samples, and these breakdowns appeared to have an effect on the scores raters awarded. It would be desirable to control for this and only include samples without breakdowns to reduce the variance due to incomprehension. The perceived difficulty of the test-tasks (even though these were validated and found to exert no effect on scores in Nakatsuhara et al., 2021a) may have caused raters to give the benefit of the doubt for more difficult tasks, or to focus more on the lack of comprehension rather than language production. Each of these factors could then skew the effects of nonverbal behavior on language scores. More comparable samples would be desirable for future research.

The design of the rating instrument was also a limiting factor in this study. Having raters judge language and affect simultaneously likely led to a halo effect across rating categories. This certainly appeared to be the case in the correlation tables presented in Chapter 5. A study design where raters scored language on one occasion and affective states on a separate occasion may have resulted in size differences

in the correlations. However, because the qualitative data triangulated with these findings, I do believe that the associations were not completely artificial. Another limitation related to the scale was the lack of definitions and more extensive practice when assigning scores for language. In the stimulated verbal recall sessions, raters quite often applied the broad definition of fluency (Lennon, 1990)—that of overall language ability—when scoring this category. Thus, fluency served as a "catch all" category that may have consumed interesting variance attributable to the narrow definition of fluency and overlapped excessively with grammar and vocabulary. The strong correlations between these three categories suggests this was the case. Comprehensibility, however, appeared to have variance distinct from that of proficiency. For future research, better benchmarked samples, more extensive practice, and specific definitions of the language categories could offer important insight on any differential effects across these categories.

Another limitation with the rating design was the choice to have raters conduct the study remotely. This made it impossible for me to control for distractions in their environment and to make sure they paid attention to the screen during the rating. It also made it impossible to ensure that multiple people were not involved in the ratings, that the participants were alert and attentive, or other concerns. However, this methodological decision was partly made because of health concerns at the time data were collected in early 2022. Although labs were able to open at that time, many students were cautious and avoided physical campus spaces. Our university was still experiencing short, temporary shutdowns at that time, when surges in COVID-19 infections occurred. Hosting the rating instrument online avoided health concerns and made the study accessible to far more participants than would have been possible with an in-person study. I put measures into place to reduce the above limitations as much as possible. I included scales of affect to essentially force raters to watch the video, as emotion and affect is largely detected through nonverbal behavior. I wrote Java code in Qualtrics so that videos would be viewed in a large format, could not be paused, and could only be seen once, again encouraging raters to attend to the videos while they could. Instructions for the study were also complete and repeated throughout the study (when signing up, when being provided the link, in the study itself, etc.), and thus the participants were well aware of my expectations.

Regarding indices extracted from iMotions, although the measures used were objective and computer-derived, there is some doubt as to the reliability and veridicality of the measures. Though studies support generally high reliability of iMotions for use in the social sciences (e.g., Dupré et al., 2020; Flynn et al., 2020; Kulke et al., 2020; Stöckli et al., 2018), if the training sets did not include individuals from a range of backgrounds, cultures, and contexts, the results may be biased. Likewise, the features that factored into engagement, valence, and attention, while somewhat documented, are each an amalgamation of various features. These clusters may then mask the effects of individual behaviors. It is likely that more salient behaviors in an online context, such as gaze aversion and smiling, would have a greater effect than their clustered counterparts. However, this can only be explored with more nuanced correlational studies of discrete behaviors.

Finally, there are always limitations to the use of stimulated recall in mixed methods designs. Although the method purports to look into raters' memories of their cognitive processes, there is a recency effect that may impair raters from truly accessing those memories. This was certainly the case with one participant, who despite providing sufficient recalls, continuously reported that they "didn't remember anything" about their ratings. Observations can be contaminated by new observations in the second viewing, thus calling into question the veridicality of the reports. However, I implemented a procedure to ensure that raters had seen videos within 24 hours of the study in order to strengthen their memories of rating the samples. I conducted the stimulated recall sessions in person, and piloted the instructions and prompts, and thus raters were aware of the focus on memories during the rating process. As detailed in Chapter 7, I also concluded that the participants had not been exposed to the research questions ahead of time as they did not focus excessively on the topic of nonverbal behavior during their sessions. Thus, I am reasonably confident that despite the limitations, the stimulated recall method provided insights that represented the rater participants' true rating processes.

Future research

There are a number of directions for future research. For one, the study design could be extended to work with a larger number of samples with a more diverse background of test takers, as well as a more

diverse pool of interlocutors and raters. The current design was restricted to Chinese test takers, mostly British interlocutors, and American raters, and as such the findings do not necessarily generalize well to other groups without further analysis. In particular, though, a study with a larger number of test taker samples would be most beneficial to examine a greater range of performances at a wider range of score levels. Likewise, different modalities, including dyadic tests in a Zoom setting, could be an interesting format to explore these effects further, as well as the effects' relationships to interactional competence.

Another area of future research is to extend the current study with data that I have already collected. The ELAN files I annotated in this study were used to produce illustrative examples for the 10 samples used in the stimulated recall design. However, all 30 of the files were annotated as part of the work with the research assistants I hired. Work from Kim et al. (2023), Trofimovich et al. (2021), and Tsunemoto et al. (2022) could be extended to include these human annotated statistics in models of the four language outcomes. This would be similar, for example, to Tsunemoto et al.'s (2022) design, in which the outcome variables were fluency, accentedness, and comprehensibility. In contrast to these studies, however, I have a dataset that is much richer, including not only frequency counts, but also duration, as well as a larger number of rated outcomes. Using a wider range of observed phenomena may reveal contrasting findings.

Likewise, using the full dataset of annotated transcripts presents an opportunity to investigate various phenomena occurring alongside talk in interaction. For example, McDonough et al. (2019, 2023) analyzed instances of dyadic communication for various features of nonverbal behavior that correlated with instances of nonunderstanding. I also conducted a small-scale multimodal conversation analytic study of four sequences of nonunderstanding, finding idiosyncratic patterns of behavior that characterized the onset of repair sequences and resolution (Burton, 2021a). With a dataset as rich as this one, this phenomenon and others could be investigated in a robust ethnomethodological design, as there are 30 transcripts of individuals with a span of proficiency levels.

Another possibility using existing data would be a methodological analysis of the accuracy and interpretability of iMotions for applied linguistics research. The iMotions dataset that was produced ultimately contained 34 variables (the three reported in this study, base emotions, and discrete behavioral

indices) and 51 Cartesian coordinate variables that represented points on the face. It may be possible to run correlations or side-by-side analyses of the ELAN data and iMotions data to support inferences from iMotions. A study of this type may look at the different issues to consider when extracting iMotions variables, including benchmarking and thresholding.

In terms of interlocutors, one of the outstanding questions in speaking assessment is the impact that examiners have on the speaking test performances of test takers. Past research has found that examiners can have a substantial impact on the language that test takers produce, resulting in score results that provide conflicting interpretations of their ability (Brown, 2003; see also Thompson et al., 2016). The examiner's affect and nonverbal behavior are also salient to test takers and can also impact the test taker's experience, making them feel more relaxed and possibly altering the examiner-candidate power dynamic (Briegel-Jones, 2014; Plough & Bogart, 2008). No research exists, to my knowledge, that has analyzed the score impact of examiners with different affective stances. Through observations of my own dataset, I noticed that when examiners smiled, the test taker often returned the smile. This type of behavior matching, alignment, or perhaps affective contagion has been found in verbal language and hypothesized in L2 nonverbal behavior (Pickering & Garrod, 2004). Recent research has found that when behavioral alignment occurs in L2 speaking dyads, it predicts increases in participant motivation (McDonough et al., 2022a). If this is indeed the case, rater behavior may impact not only the comfort and power dynamics in an oral proficiency test, but also the performances and resulting scores. By leveraging automated analyses of behavior, this type of study may be more feasible to carry out than before.

Finally, there are a range of other ideas to investigate in the future. After more research is conducted into the behaviors that appear at different stages of fluency or vocabulary development, scales could be constructed, trialed, and compared with language-only scales to determine whether adding nonverbal behavior and affect to these scales is meaningful and effective. Some research in this area has already been conducted (e.g., Jenkins & Parra, 2003). Likewise, as detailed in the discussion section, models of L2 communicative competence need revision and extension. Eventually, work to include the topics included in this dissertation would be valuable for applied linguists and others that use these frameworks when

developing assessment instruments.

Final word

This dissertation represents the culmination of a large body of research on an understudied topic within the field of language testing. Though certainly not the first study of its kind, it has added substantially to our understanding of the relationship among nonverbal behavior, affect, and language proficiency. Its limitations are diverse, as is the case with all research, but despite these limitations, the triangulated findings are interpretable and generalizable to at least the populations sampled. Much more needs to be done to extend this work in various directions to confirm the findings and determine which aspects of this area are most applicable. I sincerely hope that my efforts here can make a positive outcome on language testing practice. For one, it can benefit learners by taking into account the much wider realm of visual communication beyond linguistic resources, thus recognizing the learners' full repertoires of abilities when communicating in their second language. It can also benefit raters, as without descriptors fully representing targeted test constructs, they may rate using their own set of internal criteria, drawing from the visual realm when it is not represented. Drafting rating scales that better represent the construct would also benefit score users, as these individuals would have a better representation of the test taker's abilities to communicate.

REFERENCES

Affectiva. (n. d.). *Affectiva media analytics*. https://go.affectiva.com/affdex-for-market-research

Afifi, T. D., & Denes, A. (2013). Feedback processes and physiological responding. In In J. A. Hall & M. L. Knapp (Eds.), *Nonverbal communication* (pp. 333–368). De Gruyter. https://doi.org/10.1515/9783110238150.333

Ahammer, A., Lackner, M., & Voigt, J. (2019). Does confidence enhance performance? Causal evidence from the field. *Managerial and Decision Economics, 40*(6), 704–717. https://doi.org/10.1002/mde.3038

Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, *15*(6), 593–613. https://doi.org/10.1080/016909600750040571

Allen, L. Q. (1995). The effect of emblematic gestures on the development and access of mental representations of French expressions. *Modern Language Journal*, *79*(4), 521–529. https://doi.org/10.1111/j.1540-4781.1995.tb05454.x

Ambady, N. (2010). The perils of pondering: Intuition and thin slice judgments. *Psychological Inquiry*, *21*(4), 271–278. https://doi.org/10.1080/1047840X.2010.524882

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Argyle, M. (1988). *Bodily communication* (2nd ed.). Methuen.

Argyle, M., & Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry*, *28*, 289–304. https://doi.org/10.2307/2786027

Arnold, M. B. (1960). *Emotion and personality: (Vol. 1) Psychological aspects*. Columbia University Press.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99–115. https://doi.org/10.1177/0265532215582283

Aziz, J. R., & Nicoladis, E. (2019). "My French is rusty": Proficiency and bilingual gesture use in a majority English community. *Bilingualism: Language and Cognition*, *22*(4), 826–835. https://doi.org/10.1017/S1366728918000639

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, *16*, 449–465. https://doi.org/10.2307/3586464

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language test*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Back, M. D., Schmukle, S. C., & Egloff, B. (2011). A closer look at first sight: Social relations lens model analysis of personality and interpersonal attraction at zero acquaintance. *European Journal of Personality*, *25*(3), 225–238. https://doi.org/10.1002/per.790

Baralt, M., Gurzynski-Weiss, L., & Kim, Y. (2016). Engagement with the language: How examining learners' affective and social engagement explains successful learner-generated attention to form. In M. Sato & S. Ballinger (Eds.), *Peer interaction and second language learning: Pedagogical potential and research agenda* (pp. 209–239). John Benjamins.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54–74. https://doi.org/10.1080/15434300903464418

Barnwell, D. (1989). 'Naive' native speakers and judgements of oral proficiency in Spanish. *Language Testing*, *6*(2), 152–163. https://doi.org/10.1177/026553228900600203

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Macmillan.

Batty, A. O. (2021). An eye-tracking study of attention to visual cues in L2 listening tests. *Language Testing*, *38*(4), 511–535. https://doi.org/10.1177/0265532220951504

Bavelas, J. B., Black, A., Lemery, C. R., & Mullett, J. (1986). "I *show* how you feel": Motor mimicry as a communicative act. *Journal of Personality and Social Psychology*, *50*(2), 322–329. https://doi.org/10.1037/0022-3514.50.2.322

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of communication*, *52*(3), 566–580. https://doi.org/10.1111/j.1460-2466.2002.tb02562.x

Beattie, G., & Shovelton, H. (1999). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology*, *18*(4), 438–462. https://doi.org/10.1177/0261927X99018004005

Belío-Apaolaza, H. S., & Hernández Muñoz, N. (2021). Emblematic gestures learning in Spanish as L2/FL: Interactions between types of gestures and tasks. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688211006880

Beltrán, J. (2016). The Effects of visual input on scoring a speaking achievement test. *Working Papers in TESOL & Applied Linguistics*, *16*(2), 1–24. https://doi.org/10.7916/D8795GKM

Benazzo, S., & Morgenstern, A. (2014). A bilingual child's multimodal path into negation. *Gesture*, *14*(2), 171–202. https://doi.org/10.1075/gest.14.2.03ben

Berry, V. (2007). *Personality differences and oral test performance*. Peter Lang.

Birdwhistle, R. (1970). *Kinesics and context*. University of Pennsylvania Press. https://doi.org/10.9783/9780812201284

Blairy, S., Herrera, P., & Hess, U. (1999). Mimicry and the judgement of emotional facial expressions. *Journal of Nonverbal Behavior, 23*, 5–41. https://doi.org/10.1023/A:1021370825283

Bloom, B. (1953). Thought-processes in lectures and discussions. *Journal of General Education*, *7*(3), 160–169. https://www.jstor.org/stable/27795429

Boiger, M., & Mesquita, B. (2012). The construction of emotion in interactions, relationships, and cultures. *Emotion Rreview*, *4*(3), 221–229. https://doi.org/10.1177/1754073912439765

Boiten, F. (1996). Autonomic response patterns during voluntary facial action. *Psychophysiology*, *33*(2), 123–131. https://doi.org/10.1111/j.1469-8986.1996.tb02116.x

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge. https://doi.org/10.4324/9780429030499

Borkenau, P., Brecke, S., Möttig, C., & Paelecke, M. (2009). Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, *43*(4), 703–706. https://doi.org/10.1016/j.jrp.2009.03.007

Botes, E., Dewaele, J.-M., & Greiff, S. (2020). The power to improve: Effects of multilingualism and perceived proficiency on enjoyment and anxiety in foreign language learning. *European Journal of Applied Linguistics*, *8*(2), 1–28. http://doi.org/10.1515/eujal-2020-0003

Bourdieu, P. (1977). *Outline of a theory of practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511812507

Bourgeois, P., & Hess, U. (2008). The impact of social context on mimicry. *Biological Psychology*, *77*(3), 343–352. https://doi.org/10.1016/j.biopsycho.2007.11.008

Bowles, M. (2018). Introspective verbal reports: Think-alouds and stimulated recall. In A. Phakiti, P. De Costa, L. Plonsky & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 339–357). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-59900-1

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, *46*(4), 1171–1178. https://doi.org/10.2307/2532457

Briegel-Jones, L. (2014). *An investigation into the nonverbal behavior in the oral proficiency interview: Perceptions of interview variability and the impact on candidates* [Unpublished MA thesis]. Newcastle University, United Kingdom.

Briner, R. B., & Kiefer, T. (2005). Psychological research into the experience of emotion at work: Definitely older, but are we any wiser? In N. M. Ashkanasy, C. E. J. Hartel, & W. J. Zerbe (Eds.), *Research on emotion in organizations: The effect of affect in organizational settings* (pp. 281–307). Emerald Group Publishing. https://doi.org/10.1016/S1746-9791(05)01112-0

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, *20*(1), 1–25. https://doi.org/10.1191/0265532203lt242oa

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Research Report Series*, *2005*(1), i–157. https://doi.org/10.1002/j.2333-8504.2005.tb01982.x

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press. https://doi.org/10.1017/CBO9780511813085

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press. https://doi.org/10.1525/9780520350519

Buck, R. (1984). *The Communication of Emotion*. Guilford.

Buck, R., & Powers, S. R. (2006). The biological foundations of social organization: The dynamic emergence of social structure through nonverbal communication. In V. Manusov & M. L. Patterson (Eds.), *The Sage handbook of nonverbal communication* (pp. 119–138). Sage Publications. https://doi.org/10.4135/9781412976152.n7

Buck, R., & VanLear, C. A. (2002). Verbal and nonverbal communication: Distinguishing symbolic, spontaneous, and pseudo-spontaneous nonverbal behavior. *Journal of communication*, *52*(3), 522–541. https://doi.org/10.1111/j.1460-2466.2002.tb02560.x

Burch, A. R., & Kley, K. (2020). Assessing interactional competence: The role of intersubjectivity in a paired-speaking task. *Papers in Language Testing and Assessment*, *9*(1), 25–63. http://www.altaanz.org/uploads/5/9/0/8/5908292/2020_9_1__2_burch_kley.pdf

Burgoon, J. K. (1978). A communication model of personal space violations: Explication and an initial test. *Human Communication Research*, *4*(2), 129–142. https://doi.org/10.1111/j.1468-2958.1978.tb00603.x

Burgoon, J. K., Buller, D. B., Hale, J. L., & de Turck, M. A. (1984). Relational messages associated with nonverbal behaviors. *Human Communication Research*, *10*(3), 351–378. https://doi.org/10.1111/j.1468-2958.1984.tb00023.x

Burgoon, J. K., Guerrero, L. K., & Floyd, K. (2016). *Nonverbal communication*. Routledge. https://doi.org/10.4324/9781315663425

Burton, J. D. (2020). "How scripted is this going to be?" Raters' views of authenticity in speaking-performance tests. *Language Assessment Quarterly*, *17*(3), 244–261. https://doi.org/10.1080/15434303.2020.1754829

Burton, J. D. (2021a). The face of communication breakdown: Multimodal repair in L2 oral proficiency interviews. *Papers in Language Testing and Assessment*, *10*(2), 30–61. http://www.altaanz.org/uploads/5/9/0/8/5908292/3._plta_10_2__burton.pdf

Burton, J. D. (2021b). *The impact of nonverbal behavior on second language proficiency* [Project]. https://osf.io/u6243

Burton, J. D. (2023). Gazing into cognition: Eye behavior in online L2 speaking tests. *Language Assessment Quarterly*, *23*(2), 190–214. https://doi.org/10.1080/15434303.2022.2143680

Byram, M. (2021). *Teaching and assessing intercultural communicative competence: Revisited* (2nd ed.). Multilingual Matters. https://doi.org/10.21832/9781800410251

Cacioppo, J. T., Berntson, G. G., Larsen, J. L., Poehlmann, K. M., & Ito, T. A. (2000). The physiology of emotion. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (pp. 173–191). The Guilford Press.

Cadierno, T. (2004) Expressing motion events in a second language: A cognitive typological perspective. In M. Achard & S. Niemeir (Eds.), *Cognitive linguistics, second language acquisition, and foreign language teaching* (pp. 13-43). De Gruyter. https://doi.org/10.1515/9783110199857.13

Campbell, J., Quincy, C., Osserman, J., & Pedersen, O. (2013). Coding in-depth semi-structured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research, 42*(3), 294–320. https://doi.org/10.1177/0049124113500475

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly, 3*(3), 229–242. https://doi.org/10.1207/s15434311laq0303_1

Canale, M. (1983). From communicative competence to communicative performance. In Richards, J. C. & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–27). Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47. https://doi.org/10.1093/applin/I.1.1

Cappella, J. N., & Greene, J. O. (1982). A discrepancy-arousal explanation of mutual influence in expressive behavior for adult and infant-adult interaction. *Communications Monographs, 49*(2), 89–114. https://doi.org/10.1080/03637758209376074

Cassell, J., McNeill, D., & McCullough, K. E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition, 7*(1), 1–33. https://doi.org/10.1075/pc.7.1.03cas

Celce-Murcia, M. (1995). The elaboration of sociolinguistic competence: Implications for teacher education. In J. E. Alatis, C. A. Straehle, & M. Ronkin (Eds.), *Linguistics and the education of language teachers: Ethnolinguistic, psycholinguistic, and sociolinguistic aspects* (pp. 699–710). Georgetown University Press.

Celce-Murcia, M. (2007). Rethinking the role of communicative competence in language teaching. In E. Alcón Soler & M. P. Safont Jordà (Eds.), *Intercultural language use and language learning* (pp. 41–57). Springer. https://doi.org/10.1007/978-1-4020-5639-0_3

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). A pedagogical framework for communicative competence: A Pedagogically motivated model with content. *Issues in Applied Linguistics, 6*(2), 5–35. https://doi.org/10.5070/L462005216

Chartrand, T. L., & Lakin, J. L. (2013). The antecedents and consequences of human behavioral mimicry. *Annual Review of Psychology, 64*, 285–308. https://doi.org/10.1146/annurev-psych-113011-143754

Choi, J. S. (2022). *Investigating test delivery modes within video-conferenced English speaking proficiency assessment* [Unpublished doctoral dissertation]. Michigan State University.

Choi, S., & Bowerman, M. (1991). Learning to express motion events in English and Korean: The influence of language-specific lexicalization patterns. *Cognition*, *41*(1–3), 83–121. https://doi.org/10.1016/0010-0277(91)90033-Z

Choi, S., & Lantolf, J. P. (2008). The representation and embodiment of meaning in L2 communication. Motion events in the speech and gesture of advanced L2 Korean and L2 English speakers. *Studies in Second Language Acquisition*, *30*(2), 191–224. https://doi.org/10.1017/S0272263108080315

Chomsky, N. (1965). *Aspects of the theory of syntax.* MIT Press. https://doi.org/10.21236/AD0616323

Chong, J. J. Q., & Aryadoust, V. (2022). Investigating the effect of multimodality and sentiments on speaking assessments: A facial emotional analysis. *Education and Information Technologies*, *28*, 7413–7436. https://doi.org/10.1007/s10639-022-11478-7

Cienki, A. J. (2012). Usage events of spoken language and the symbolic units we (may) abstract from them. In J. Badio & K. Kosecki (Eds.), *Cognitive processes in language* (pp. 149–158). Peter Lang.

Clark, H. H. (1996). *Using language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620539

Clark, H. H. (2002). Speaking in time. *Speech communication*, *36*(1-2), 5–13. https://doi.org/10.1016/S0167-6393(01)00022-X

Clément, R. (1980). Ethnicity, contact and communicative competence in a second language. In H. Giles & W. P. Robinson (Eds.), *Language: Social psychological perspectives* (pp. 147–154). Pergamon. https://doi.org/10.1016/B978-0-08-024696-3.50027-2

Clément, R. (1986). Second language proficiency and acculturation: An investigation of the effects of language status and individual characteristics. *Journal of Language and Social Psychology*, *5*(4), 271–290. https://doi.org/10.1177/0261927X8600500403

Clément, R., & Kruidenier, B. (1983). Orientations in second language acquisition: I. The effects of ethnicity, milieu and target language on their emergence. *Language Learning*, *33*(3), 273–291. https://doi.org/10.1111/j.1467-1770.1983.tb00542.x

Clément, R., & Kruidenier, B. G. (1985). Aptitude, attitude and motivation in second language proficiency: A test of Clément's model. *Journal of language and Social Psychology*, *4*(1), 21–37. https://doi.org/10.1177/0261927X8500400102

Clément, R., Dörnyei, Z., & Noels, K. (1994). Motivation, self- confidence, and group cohesion in the foreign language classroom. *Language Learning*, *44*(3), 417–448. https://doi.org/10.1111/j.1467-1770.1994.tb01113.x

Cobb-Clark, D. A. (2015). Locus of control and the labor market. *IZA Journal of Labor Economics*, *4*, Article 3. https://doi.org/10.1186/s40172-014-0017-x

Cohen, R. L., & Otterbein, N. (1992). The mnemonic effect of speech gestures: Pantomimic and non-pantomimic gestures compared. *European Journal of Cognitive Psychology*, *4*(2), 113–139. https://doi.org/10.1080/09541449208406246

Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System*, *29*(1), 1–14. https://doi.org/10.1016/S0346-251X(00)00057-9

Conlan, C. J., Bardsley, W. N., & Martinson, S. H. (1994). *Study of intra-rater reliability of assessments of live versus audio-recorded interviews in the IELTS Speaking component* [Unpublished study]. International Editing Committee of IELTS.

Cook, S. W., Yip, T. K., & Goldin-Meadow, S. (2012). Gestures, but not meaningless movements, lighten working memory load when explaining math. *Language and Cognitive Processes*, *27*(4), 594–610. https://doi.org/10.1080/01690965.2011.567074

Cope, B., & Kalantzis, M. (2020). *Making sense: Reference, agency, and structure in a grammar of multimodal meaning*. Cambridge University Press. https://doi.org/10.1017/9781316459645

Corbin, J., & Strauss, A. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (4th ed.). Sage.

Coulson, M. (2004). Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, *28*, 117–139. https://doi.org/10.1023/B:JONB.0000023655.25550.be

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume.* Council of Europe. https://rm.coe.int/cefr-companion-volume-with-new-descriptors- 2018/1680787989

Cox, T. L., Brown, A. V., & Thompson, G. L. (2022). Temporal fluency and floor/ceiling scoring of intermediate and advanced speech on the ACTFL Spanish Oral Proficiency Interview–computer. *Language Testing*. Advance online publication. https://doi.org/10.1177/02655322221114614

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. Sage publications.

Crivelli, C., & Fridlund, A. J. (2018). Facial displays are tools for social influence. *Trends in Cognitive Sciences*, *22*(5), 388–399. https://doi.org/10.1016/j.tics.2018.02.006

Crivelli, C., Jarillo, S., Russell, J. A., & Fernández-Dols, J. M. (2016). Reading emotions from faces in two indigenous societies. *Journal of Experimental Psychology: General*, *145*(7), 830–843. https://psycnet.apa.org/doi/10.1037/xge0000172

Crowther, D., Holden, D., & Urada, K. (2022). Second language speech comprehensibility. *Language Teaching*, *55*(4), 470–489. https://doi.org/10.1017/S0261444821000537

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92*(4), 631–648. https://doi.org/10.1037/0022-3514.92.4.631

Cuddy, A. J. C., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior*, *31*, 73–98. https://doi.org/10.1016/j.riob.2011.10.004

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology*, *40*, 61–149. https://doi.org/10.1016/S0065-2601(07)00002-0

Cuddy, A. J., Wilmuth, C. A., Yap, A. J., & Carney, D. R. (2015). Preparatory power posing affects nonverbal presence and job interview performance. *Journal of Applied Psychology*, *100*(4), 1286–1295. https://doi.org/10.1037/a0038543

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31–51. https://doi.org/10.1177/026553229000700104

Cutrone, P. (2005). A case study examining backchannels in conversations between Japanese–British dyads. *Multilingua*, *24*(3), 237–274. https://doi.org/10.1515/mult.2005.24.3.237

Dael, N., Mortillaro, M., & Scherer, K. R. (2012). Emotion expression in body action and posture. *Emotion, 12*(5), 1085–1101. https://doi.org/10.1037/a0025737

Dahl, T. I., & Ludvigsen, S. (2014). How I see what you're saying: The role of gestures in native and foreign language listening comprehension. *The Modern Language Journal*, *98*(3), 813–833. https://doi.org/10.1111/modl.12124

Dai, D. W. (2023). What do second language speakers really need for real-world interaction? A needs analysis of L2 Chinese interactional competence. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688221144836

Dao, P., & McDonough, K. (2018). Effect of proficiency on Vietnamese EFL learners' engagement in peer interaction. *International Journal of Educational Research, 88*, 60–72. https://doi.org/10.1016/j.ijer.2018.01.008

Davies, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367–396. https://doi.org/10.1177/0265532209104667

De Costa, P. I. (2016). *The power of identity and ideology in language learning: Designer immigrants learning English in Singapore*. Springer.

de Jong, N. H. (2023). Assessing second language speaking proficiency. *Annual Review of Linguistics, 9*, 541–560. https://doi.org/10.1146/annurev-linguistics-030521-052114

De Rivera, J., & Grinkis, C. (1986). Emotions as social relationships. *Motivation and emotion, 10*, 351-369. https://doi.org/10.1007/BF00992109

De Rivera, J., & Grinkis, C. (1986). Emotions as social relationships. *Motivation and Emotion, 10*, 351–369. https://doi.org/10.1007/BF00992109

De Riviera, J. (1977). *A structural theory of emotions*. International Universities Press.

DeKeyser, R. (2020). Skill acquisition theory. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition* (pp. 83-104). Routledge. https://doi.org/10.4324/9780429503986-5

DeKeyser, R. M. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, *19*(2), 195–221. https://doi.org/10.1017/S0272263197002040

DePaulo, B. M., & Friedman, H. S. (1998). Nonverbal communication. In S. T. Fiske, D. Gilbert, & G. Lindzey (Eds.), *The handbook of social psychology* (3rd ed., Vol. 2, pp. 3–40). McGraw-Hill.

Derrida, J. (1967). *De la grammatologie*. Les Éditions de Minuit.

Dewaele, J.-M. (2010). Multilingualism and affordances: Variation in self-perceived communicative competence and communicative anxiety in French L1, L2, L3 and L4. *International Review of Applied Linguistics*, *48*(2–3), 105–129. https://doi.org/10.1515/iral.2010.006

Dewaele, J.-M., & Alfawzan, M. (2018). Does the effect of enjoyment outweigh that of anxiety in foreign language performance? *Studies in Second Language Learning and Teaching*, *8*(1), 21–45. https://doi.org/10.14746/ssllt.2018.8.1.2

Dewaele, J.-M., & Li, C. (2020). Emotions in second language acquisition: A critical review and research agenda. *Foreign Language World*, *196*(1), 34-49.

Dewaele, J.-M., & Li, C. (2022). Foreign language enjoyment and anxiety: Associations with general and domain specific English achievement. *Chinese Journal of Applied Linguistics*, *45*(1), 23–48. https://doi.org/10.1515/cjal-2022-0104

Dewaele, J.-M., & MacIntyre, P. D. (2014). The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching*, *4*(2), 237–274. http://doi.org/ 10.14746/ssllt.2014.4.2.5

Dewaele, J.-M., Özdemir, C., Karci, D., Uysal, S., Özdemir, E. D., & Balta, N. (2019). How distinctive is the foreign language enjoyment and foreign language classroom anxiety of Kazakh learners of Turkish? *Applied Linguistics Review*, *13*(2), 243–265. https://doi.org/10.1515/applirev-2019-0021

Diener, E., Smith, H., & Fujita, F. (1995). The personality structure of affect. *Journal of Personality and Social Psychology*, *69*(1), 130–141. https://doi.org/10.1037/0022-3514.69.1.130

Dimberg, U., Thunberg, M., & Elmehed, K. (2000). Unconscious facial reactions to emotional facial expressions. *Psychological science*, *11*(1), 86–89. https://doi.org/10.1111/1467-9280.00221

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., Rossi, G., & Enfield, N. J. (2015). Universal principles in the repair of communication problems. *PloS one*, *10*(9), Article e0136100. https://doi.org/10.1371/journal.pone.0136100

Doherty-Sneddon, G., & Phelps, F. G. (2005). Gaze aversion: A response to cognitive or social difficulty? *Memory and Cognition*, *33*(4), 727–733. https://doi.org/10.3758/bf03195338

Doherty-Sneddon, G., Bruce, V., Bonner, L., Longbotham, S., & Doyle, C. (2002). Development of gaze aversion as disengagement from visual information. *Developmental Psychology*, *38*(3), 438–445. https://doi.org/10.1037//0012-1649.38.3.438

Doqaruni, V. (2015). Increasing confidence to decrease reticence: A qualitative action research in second language education. *Canadian Journal of Action Research, 16*(3), 42–60. https://doi.org/10.33524/cjar.v16i3.227

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*(2), 125–144. https://doi.org/10.1177/026553229401100203

Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research, 60*(1), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Ducasse, A. (2010). *Interaction in paired oral proficiency assessment in Spanish*. Peter Lang. https://doi.org/10.3726/978-3-653-05393-7

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423–443. https://doi.org/10.1177/0265532209104669

Duncan, S., & Fiske, D. W. (2015). *Face-to-face interaction: Research, methods, and theory*. Routledge. https://doi.org/10.4324/9781315660998

Dupré, D., Krumhuber, E. G., Küster, D., & McKeown, G. J. (2020). A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PLoS ONE, 15*(4), Article e0231968. https://doi.org/10.1371/journal.pone.0231968

Educational Testing Services (ETS). (n.d.). *TOEFL iBT integrated speaking rubrics*. Educational Testing Services. https://www.ets.org/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf

Edwards, E., & Roger, P. S. (2015). Seeking out challenges to develop L2 self-confidence: A language learner's journey to proficiency. *TESL-EJ, 18*(4), 1–24. http://tesl-ej.org/pdf/ej72/a3.pdf

Egi, T. (2008). Investigating stimulated recall as a cognitive measure: Reactivity and verbal reports in SLA research methodology. *Language Awareness, 17*(3), 212–228. https://doi.org/10.1080/09658410802146859

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), *Nebraska symposium on motivation* (pp. 207–283). University of Nebraska Press.

Ekman, P., & Friesen, W. V. (1974). Nonverbal behavior and psychopathology. In R. J. Friedman & M. M. Katz (Eds.), *The psychology of depression: Contemporary theory and research* (pp. 203–232). John Wiley & Sons.

Ekman, P., & Frisen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1*(1), 49–98. https://doi.org/10.1515/9783110880021.57

Ekman, P., & Rosenberg, E. (Eds.) (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)* (2nd ed). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195179644.001.0001

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial action coding system*. Research Nexus, Network Research Information.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, *164*(3875), 86–88. https://doi.org/10.1126/science.164.3875.86

Elfenbein, H. A. (2014). The many faces of emotional contagion: An affective process theory of affective linkage. *Organizational Psychology Review*, *4*(4), 326–362. https://doi.org/10.1177/2041386614542889

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, *128*(2), 203–235. https://doi.org/10.1037/0033-2909.128.2.203

Elfenbein, H. A., Beaupré, M., Lévesque, M., & Hess, U. (2007). Toward a dialect theory: cultural differences in the expression and recognition of posed facial expressions. *Emotion*, *7*(1), 131–146. https://doi.org/10.1037/1528-3542.7.1.131

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford University Press.

Engle, R. A. (1998). Not channels but composite signals: Speech, gesture, diagrams and object demonstrations are integrated in multimodal explanations. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the twentieth annual conference of the cognitive science society* (pp. 321–326). Routledge. https://doi.org/10.4324/9781315782416-65

Ericsson, K.A, & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

FaceReader. (n.d.). *Noldus FaceReader*. https://www.noldus.com/facereader

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: a narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in psychology*, *11*, Article 330. https://doi.org/10.3389/fpsyg.2020.00330

Firth, A., & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *Modern Language Journal, 81*(3), 285–300. https://doi.org/10.1111/j.1540-4781.2007.00667.x

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Floyd, S., Manrique, E., Rossi, G., & Torreira, F. (2016). Timing of visual bodily behavior in repair sequences: Evidence from three languages. *Discourse Processes*, *53*(3), 175–204. https://doi.org/10.1080/0163853X.2014.992680

Flynn, M., Effraimidis, D., Angelopoulou, A., Kapatanios, E., Williams, D., Hemanth, J., & Towell, T. (2020). Assessing the effectiveness of automated emotion recognition in adults and children for clinical investigation. *Frontiers in Human Neuroscience*, *14*(70), Article 70. https://doi.org/10.3389/fnhum.2020.00070

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist, 56*(3), 218–226. https://doi.org/10.1037/0003-066X.56.3.218

Fredrickson, B. L. (2003). The value of positive emotions: The emerging science of positive psychology is coming to understand why it's good to feel good. *American Scientist, 91*(4), 330–335. https://doi.org/10.1511/2003.26.330

Frick-Horbury, D., & Guttentag, R. E. (1998). The effects of restricting hand gesture production on lexical retrieval and free recall. *The American Journal of Psychology, 111*(1), 43–62. https://doi.org/10.2307/1423536

Fridlund, A. J. (1994). *Human facial expression: An evolutionary view.* Academic Press.

Frijda, N. (2005). Emotion experience. *Cognition & Emotion, 19*(4), 473–497. https://doi.org/10.1080/02699930441000346

Frijda, N. H. (1986). *The emotions*. Cambridge University Press.

Frijda, N. H. (1994). Varieties of affect: Emotions and episodes, moods, and sentiments. In R. J. Davidson (Ed.), *The nature of emotion – fundamental questions* (pp. 59–67). Oxford University Press.

Frijda, N. H., & Mesquita, B. (1994). The social roles and functions of emotions. In S. Kitayama & H. R. Markus (Eds.), *Emotion and culture: Empirical studies of mutual influence* (pp. 51–87). American Psychological Association. https://doi.org/10.1037/10152-002

Fulcher, G. (2003). *Testing second language speaking*. Pearson Longman.

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Gan, Z., & Davison, C. (2011). Gestural behavior in group oral assessment: A case study of higher- and lower-scoring students. *International Journal of Applied Linguistics, 21*(1), 95–120. https://doi.org/10.1111/j.1473-4192.2010.00264.x

Gardner, R. C., & MacIntyre, P. D. (1993). On the measurement of affective variables in second language learning. *Language Learning, 43*(2), 157–194. https://doi.org/10.1111/j.1467-1770.1992.tb00714.x

Gass, S. M., & Mackey, A. (2016). *Stimulated recall methodology in applied linguistics and L2 research* (2nd ed.). Routledge. https://doi.org/10.4324/9781315813349

Gass, S. M., & Mackey, A. (2020). Input, interaction, and output in L2 Acquisition. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition* (pp. 192–222). Routledge. https://doi.org/10.4324/9780429503986-9

Gifford, R. (2013). Personality is encoded in, and decoded from, nonverbal behavior. In J. A. Hall & M. K. Knapp (Eds.), *Nonverbal communication* (pp. 369–402). De Gruyter Mouton. https://doi.org/10.1515/9783110238150.369

Givens, D. B., & White, J. (2021). *The Routledge dictionary of nonverbal communication*. Routledge. https://doi.org/10.4324/9780429293665

Gkonou, C., Daubney, M., & Dewaele, J.-M. (Eds.). (2017). *New insights into language anxiety: Theory, research, and educational implications*. Multilingual Matters. https://doi.org/10.21832/9781783097722

Glenberg, A. M., Schroeder, J. L., & Robertson, D. A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition*, *26*(4), 651–658. https://doi.org/10.3758/BF03211385

Godfroid, A. (2019). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. Routledge. https://doi.org/10.4324/9781315775616

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. The Belknap Press. https://doi.org/10.1037/e413812005-377

Goldin-Meadow, S., & Alibali, M. W. (2013). Gesture's role in speaking, learning, and creating language. *Annual Review of Psychology*, *64*, 257–283. https://doi.org/10.1146/annurev-psych-113011-143802

Goldin-Meadow, S., & Brentari, D. (2017). Gesture, sign, and language: The coming of age of sign language and gesture studies. *Behavioral and Brain Sciences*, *40*, Article E46. https://doi.org/10.1017/S0140525X15001247

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12*(6), 516–522. https://doi.org/10.1111/1467-9280.00395

Goldin-Meadow, S., Wein, D., & Chang, C. (1992). Assessing knowledge through gesture: Using children's hands to read their minds. *Cognition and Instruction*, *9*(3), 201–219. https://doi.org/10.1207/s1532690xci0903_2

Goodwin, C. (1980). Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning. *Sociological Inquiry*, *50*(3-4), 272–302. https://doi.org/10.1111/j.1475-682X.1980.tb00023.x

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, *32*(10), 1489–1522. https://doi.org/10.1016/S0378-2166(99)00096-X

Goodwin, C. (2018). *Co-operative action*. Cambridge University Press. https://doi.org/10.1017/9781139016735

Goturk, N., & Chukharev-Hudilainen, E. (2023). Strategy use in a spoken dialog system-delivered paired discussion task: A stimulated recall study. *Language Testing*. Advance online publication. https://doi.org/10.1177/02655322231152620

Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology*, *10*(1), 57–67. https://doi.org/10.1080/00207597508247319

Graham, J. A., & Heywood, S. (1975). The effects of elimination of hand gestures and of verbal codability on speech performance. *European Journal of Social Psychology*, *5*(2), 189–195. https://doi.org/10.1002/ejsp.2420050204

Gray, H. M., & Ambady, N. (2006). Methods for the study of nonverbal communication. In V. Manusov & M. L. Patterson (Eds.), *The Sage handbook of nonverbal communication* (pp. 41–58). Sage Publications. https://doi.org/10.4135/9781412976152.n3

Graziano, M., & Gullberg, M. (2018). When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in Psychology*, *9*, Article 879. https://doi.org/10.3389/fpsyg.2018.00879

Greer, T., & Potter, H. (2008). Turn-taking practices in multi-party EFL oral proficiency tests. *Journal of Applied Linguistics*, *5*(3), 295–318. https://doi.org/10.1558/japl.v5i3.297

Gregersen, T. S. (2005). Nonverbal cues: Clues to the detection of foreign language anxiety. *Foreign Language Annals*, *38*(3), 388-400. https://doi.org/10.1111/j.1944-9720.2005.tb02225.x

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

Gregersen, T., Olivares-Cuhat, G., & Storm, J. (2009). An examination of L1 and L2 gesture use: What role does proficiency play? *Modern Language Journal*, *93*(2), 195–208. https://doi.org/10.1111/j.1540-4781.2009.00856.x

Groeber, S., & Pochon-Berger, E. (2014). Turns and turn-taking in sign language interaction: A study of turn-final holds. *Journal of Pragmatics*, *65*, 121–136. https://doi.org/10.1016/j.pragma.2013.08.012

Guerin, B. (1986). Mere presence effects in humans: A review. *Journal of Experimental Social Psychology*, *22*(1), 38–77. https://doi.org/10.1016/0022-1031(86)90040-5

Guerrero, L. K., & Wiedmaier, B. (2013). Nonverbal intimacy: Affectionate communication, positive involvement behavior, and flirtation. In J. A. Hall & M. L. Knapp (Eds.), *Nonverbal communication* (pp. 577–612). De Gruyter. https://doi.org/10.1515/9783110238150.577

Gullberg, M. (1998). *Gesture as a communication strategy in second language discourse: A study of learners of French and Swedish*. Lund University Press. https://lup.lub.lu.se/search/files/4825091/3912717.pdf

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *IRAL - International Review of Applied Linguistics in Language Teaching*, *44*(2), 103–124. https://doi.org/10.1515/IRAL.2006.004

Gullberg, M. (2009a). Gestures and the development of semantic representations in first and second language acquisition. *Acquisition et interaction en langue étrangère*, *28*(1), 117–139. https://doi.org/10.4000/aile.4514

Gullberg, M. (2009b). Reconstructing verb meaning in a second language: How English speakers of L2 Dutch talk and gesture about placement. *Annual Review of Cognitive linguistics, 7*(1), 221–244. https://doi.org/10.1075/arcl.7.09gul

Gullberg, M. (2012). Gesture analysis in second language acquisition. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Wiley. https://doi.org/10.1002/9781405198431

Gullberg, M. (2022). Bimodal convergence: How languages interact in multicompetent language users' speech and gestures. In A. Morgenstern & S. Goldin-Meadow (Eds.), *Gesture in language: Development across the lifespan* (pp. 317–333). American Psychological Association. https://doi.org/10.1037/0000269-013

Gullberg, M., & De Bot, K. (Eds.) (2010). *Gestures in language development* (Vol. 28). John Benjamins Publishing. https://doi.org/10.1075/bct.28

Halberstadt, A. G., Parker, A. E., & Castro, V. L. (2013). Nonverbal communication: Developmental perspectives. In J. A. Hall & M. L. Knapp (Eds.), *Nonverbal communication* (pp. 93–127). De Gruyter. https://doi.org/10.1515/9783110238150.93

Hall, E. T. (1963). A system for the notation of proxemic behavior. *American Anthropologist, 65*(5), 1003–1026. https://doi.org/10.1525/aa.1963.65.5.02a00020

Hall, J. A. (1984). *Nonverbal sex differences: Communication accuracy and expressive style*. The Johns Hopkins University Press. https://doi.org/10.56021/9780801824401

Hall, J. A. (2006). Women's and men's nonverbal communication. In V. Manusov & M. L. Patterson (Eds.), *The sage handbook of nonverbal communication* (pp. 201–218). Sage Publications. https://doi.org/10.4135/9781412976152.n11

Hall, J. A., & Gunnery, S. D. (2013). Gender differences in nonverbal communication. In J. A. Hall & M. K. Knapp (Eds.), *Nonverbal communication* (pp. 639–696). De Gruyter Mouton. https://doi.org/10.1515/9783110238150.639

Hall, J. A., & Knapp, M. L. (Eds.). (2013). *Nonverbal communication*. De Gruyter. https://doi.org/10.1515/9783110238150

Hall, J. A., Horgan, T. G., & Murphy, N. A. (2019). Nonverbal communication. *Annual Review of Psychology*, *70*, 271–294. https://doi.org/10.1146/annurev-psych-010418-103145

Halleck, G. B. (1992). The oral proficiency interview: Discrete point test or a measure of communicative language ability? *Foreign Language Annals, 25*(3), 227–231. https://doi.org/10.1111/j.1944-9720.1992.tb00532.x

Halliday, M. A. K. (1985). *An introduction to functional grammar*. Edward Arnold.

Harding, L. (2011). *Accent and listening assessment: A validation study of the use of speakers with L2 accents on an academic English listening test.* Peter Lang.

Harding, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly*, *11*(2), 186–197. https://doi.org/10.1080/15434303.2014.895829

Hardison, D. M. (2018). Effects of contextual and visual cues on spoken language processing: Enhancing L2 perceptual salience through focused training. In S. M. Gass, P. Spinner, & J. Behney (Eds.), *Salience in second language acquisition* (pp. 201–220). Routledge. https://doi.org/10.4324/9781315399027-11

Hardison, D., & Pennington, M. C. (2021). Multimodal second-language communication: Research findings and pedagogical implications. *RELC Journal*, *52*(1), 62–76. https://doi.org/10.1177/0033688220966635

Harrell, F. (2020, September 20). *Violation of proportional odds is not fatal*. Statistical Thinking. https://www.fharrell.com/post/po

Harrigan, J. A. (2013). Methodology: coding and studying nonverbal behavior. In In J. A. Hall & M. L. Knapp (Eds.), *Nonverbal communication* (pp. 35–68). De Gruyter. https://doi.org/10.1515/9783110238150.35

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

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). *Emotional contagion*. Cambridge University Press. https://doi.org/10.1017/CBO9781139174138

Heaver, B., & Hutton, S. B. (2011). Keeping an eye on the truth? Pupil size changes associated with recognition memory. *Memory*, *19*(4), 398–405. https://doi.org/10.1080/09658211.2011.575788

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, *24*(3), 411–482. https://doi.org/10.1086/504455

Heidari, K. (2019). Willingness to communicate: A predictor of pushing vocabulary knowledge from receptive to productive. *Journal of Psycholinguistic Research*, *48*, 903–920. https://doi.org/10.1007/s10936-019-09639-w

Hellermann, J. (2008). *Social actions for classroom language learning*. Multilingual Matters. https://doi.org/10.21832/9781847690272

Heng, C. S., Abdullah, A. N., & Yusof, N. B. (2012). Investigating the construct of anxiety in relation to speaking skills among ESL tertiary learners. *3L: The Southeast Asian Journal of English Studies*, *18*(3), 155–166.

Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In Atkinson & Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 299–345). Cambridge University Press.

Heritage, J. (1998). Oh-prefaced responses to inquiry. *Language in society*, *27*(3), 291-334. https://doi.org/10.1017/S0047404500019990

Hess, U., & Fischer, A. (2013). Emotional mimicry as social regulation. *Personality and Social Psychology Review*, *17*(2), 142–157. https://doi.org/10.1177/1088868312472607

Hirschmüller, S., Schmukle, S. C., Krause, S., Back, M. D., & Egloff, B. (2018). Accuracy of self-esteem judgments at zero acquaintance. *Journal of Personality*, *86*(2), 308–319. https://doi.org/10.1111/jopy.12316

Hirt, F., Werlen, E., Moser, I., & Bergamin, P. (2019). Measuring emotions during learning: lack of coherence between automated facial emotion recognition and emotional experience. *Open Computer Science*, *9*(1), 308–317. https://doi.org/10.1515/comp-2019-0020.

Hırçın Çoban, M., & Sert, O. (2020). Resolving interactional troubles and maintaining progressivity in paired speaking assessment in an EFL context. *Papers in Language Testing and Assessment*, *9*(1), 64–94. http://www.altaanz.org/uploads/5/9/0/8/5908292/2020_9_1__3_hircin-soban_sert.pdf

Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic Bulletin and Review*, *25*(5), 1900–1908. https://doi.org/10.3758/s13423-017-1363-z

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

Hox, J., Moerbeck, M., & van de Schoot, R. (2018). *Multilevel analysis: techniques and applications* (3rd ed.). Routledge. https://doi.org/10.4324/9781315650982

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins. https://doi.org/10.1075/lllt.41

Hulstijn, J. H., Young, R. F., Ortega, L., Bigelow, M., DeKeyser, R., Ellis, N. C., Lantolf, J. P., Mackey, A., & Talmy, S. (2014). Bridging the gap: Cognitive and social approaches to research in second language learning and teaching. *Studies in Second Language Acquisition, 36*(3), 361–421. https://doi.org/10.1017/S0272263114000035

Hymes, D. (1972). On communicative competence. In A. Duranti (Ed.), *Linguistic anthropology: A reader* (pp. 53–73). Blackwell.

iMotions. (2017). *Facial expression analysis*. https://imotions.com/biosensor/fea-facial-expression-analysis

Inceoglu, S. (2016). Effects of perceptual training on second language vowel perception and production. *Applied Psycholinguistics, 37*(5), 1175–1199. https://doi.org/10.1017/S0142716415000533

Interagency Language Roundtable. (n.d.). *Interagency Language Roundtable language skill level descriptions – Speaking.* https://www.govtilr.org/Skills/ILRscale2.htm

International English Language Testing System (IELTS). (n.d.). *Speaking: Band descriptors (public)*. IELTS. https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx

Iwaniec, J. (2019). Questionnaires: Implications for effective implementation. In J. McKinley & H. Rose (Eds.), *The Routledge handbook of research methods in applied linguistics* (pp. 324–335). Routledge. https://doi.org/10.4324/9780367824471-28

Jackson, J. C., Watts, J., Henry, T. R., List, J. M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., & Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, *366*(6472), 1517–1522. https://doi.org/10.1126/science.aaw8160

Jakobovits, L. A. (1970). *Foreign language learning.* Newbury House.

James, W. (1884). What is an emotion? *Mind*, *9*(34), 188–205. https://doi.org/10.1093/mind/os-IX.34.188

Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *The Modern Language Journal*, *87*(1), 90–107. https://doi.org/10.1111/1540-4781.00180

Jiang, Y., & Dewaele, J.-M. (2019). How unique is the foreign language classroom enjoyment and anxiety of Chinese EFL learners? *System*, *82*(59), 13–25. http://doi.org/10.1016/J.SYSTEM.2019.02.017

Jin, Y., De Bot, K., & Keijzer, M. (2017). Affective and situational correlates of foreign language proficiency: A study of Chinese university learners of English and Japanese. *Studies in Second Language Learning and Teaching*, *7*(1), 105–125. https://doi.org/10.14746/ssllt.2017.7.1.6

Judge, T. A., & Hurst, C. (2007). Capitalizing on one's advantages: Role of core self-evaluations. *Journal of Applied Psychology*, *92*(5), 1212–1227. https://doi.org/10.1037/0021-9010.92.5.1212

Jungheim, N. O. (2001). The unspoken element of communicative competence: Evaluating language learners' nonverbal behavior. In T. Hudson & J. Brown (Eds.), *A focus on language test development: Expanding the language proficiency construct across a variety of tests* (pp. 1-35). University of Hawaii, Second Language Teaching and Curriculum Center.

Kalantzis, M., & Cope, B. (2020). *Adding sense: Context and interest in a grammar of multimodal meaning*. Cambridge University Press. https://doi.org/10.1017/9781108862059

Kappas, A., Krumhuber, E., & Küster, D. (2013). Facial behavior. In J. A. Hall & M. L. Knapp (Eds.), *Nonverbal communication* (pp. 131–165). De Gruyter. https://doi.org/10.1515/9783110238150.131

Kasper, G., & Wagner, J. (2018). Epistemological reorientations and L2 interactional settings: A postscript to the special issue. *The Modern Language Journal*, *102*(S1), 82–90. https://doi.org/10.1111/modl.12463

Keevallik, L. (2014). Turn organization and bodily-vocal demonstrations. *Journal of Pragmatics*, *65*, 103–120. https://doi.org/10.1016/j.pragma.2014.01.008

Kellerman, S. (1992). 'I see what you mean': The role of kinesic behaviour in listening and implications for foreign and second language learning. *Applied Linguistics*, *13*(3), 239–258. https://doi.org/10.1093/applin/13.3.239

Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of memory and Language*, *40*(4), 577–592. https://doi.org/10.1006/jmla.1999.2634

Kelly, S. D., Manning, S. M., & Rodak, S. (2008). Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education. *Language and Linguistics Compass*, 2, 569–588. https://doi.org/10.1111/j.1749-818X.2008.00067.x

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207–227). Mouton Publishers. https://doi.org/10.1515/9783110813098.207

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press. https://doi.org/10.1017/CBO9780511807572

Kendon, A. (2007). Some topics in gesture studies. In A. Esposito, M. Bratanic, E. Keller, & M. Marinaro (Eds.), *Fundamentals of verbal and nonverbal communication and the biometric issue* (pp. 3–19). IOS Press.

Kikuchi, Y., & Noriuchi, M. (2019). Power of self-touch: its neural mechanism as a coping strategy. In S. Fukuda (Ed.), *Emotional engineering, vol. 7: The age of communication* (pp. 33–47). Springer. https://doi.org/10.1007/978-3-030-02209-9_

Kim, Y. L., Liu, C., Trofimovich, P., & McDonough, K. (2023). *Do visual cues matter for perceived fluency during L2 conversations?* [Paper presentation] American Association of Applied Linguistics. Portland, Oregon, United States.

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, *24*(2), 145–167. https://doi.org/10.1080/01690960802586188

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and language*, *48*(1), 16–32. https://doi.org/10.1016/S0749-596X(02)00505-3

Kitayama, S., Mesquita, B., & Karasawa, M. (2006). Cultural affordances and emotional experience: Socially engaging and disengaging emotions in Japan and the United States. *Journal of Personality and Social Psychology, 91*(5), 890–903. https://doi.org/10.1037/0022-3514.91.5.890

Knapp, M. L., Hall, J. A., & Horgan, T. G. (2014). *Nonverbal communication in human interaction* (8th ed.). Cengage Learning.

Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing*, *38*(4), 602–626. https://doi.org/10.1177/0265532221994052

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*(4), 366–372. https://doi.org/10.1111/j.1540-4781.1986.tb05291.x

Krason, A., Fenton, R., Varley, R., & Vigliocco, G. (2022). The role of iconic gestures and mouth movements in face-to-face communication. *Psychonomic Bulletin & Review, 29*, 600–612. https://doi.org/10.3758/s13423-021-02009-5

Krauss, R. K., Chen, Y., & Gottesman R. F. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and gesture* (pp. 261–283). Cambridge University Press. https://doi.org/10.1017/CBO9780511620850.017

Krauss, R. M., & Hadar. U. (1999). The role of speech-related arm/hand gestures in word retrieval. In R. Campbell & L. Messing (Eds.), *Language and gesture* (pp. 261–283). Cambridge University Press. https://doi.org/10.1093/acprof:oso/9780198524519.003.0006

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, *84*(3), 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing*, *31*(3), 329–348. https://doi.org/10.1177/0265532214526174

Kulke, L., Feyerabend, D., & Schacht, A. (2020). A comparison of the Affectiva iMotions facial expression analysis software with EMG for identifying facial expressions of emotion. *Frontiers in Psychology*, *11*(329), Article 329. https://doi.org/10.3389/fpsyg.2020.00329

Kunnan, A. (2018). *Evaluating language assessments*. Routledge. https://doi.org/10.4324/9780203803554

Labrie, N., & Clement, R. (1986). Ethnolinguistic vitality, self-confidence and second language proficiency: An investigation. *Journal of Multilingual & Multicultural Development, 7*(4), 269–282. https://doi.org/10.1080/01434632.1986.9994244

LaFrance, M., & Mayo, C. (1976). Racial differences in gaze behavior during conversations: Two systematic observational studies. *Journal of Personality and Social Psychology*, *33*(5), 547–552. https://doi.org/10.1037/0022-3514.33.5.547

Lakin, J. L. (2006). Automatic cognitive processes and nonverbal communication. In V. Manusov & M. L. Patterson (Eds.), *The Sage handbook of nonverbal communication* (pp. 59–77). Sage Publications. https://doi.org/10.4135/9781412976152.n4

Lam, D. M. K. (2015). Contriving authentic interaction: Task implementation and engagement in school-based speaking assessment in Hong Kong. In G. Yu & Y. Jin (Eds.), *Assessing Chinese learners of English. Language constructs, consequences and conundrums* (pp. 38–60). Palgrave Macmillan. https://doi.org/10.1057/9781137449788_3

Lam, D. M. K. (2021). Don't turn a deaf ear: A case for assessing interactive listening. *Applied Linguistics*, *42*(4), 740–764. https://doi.org/10.1093/applin/amaa064

Lambert, C., Philp, J., & Nakamura, S. (2017). Learner-generated content and engagement in second language task performance. *Language Teaching Research, 21*(6), 665–680. https://doi.org/10.1177/1362168816683559

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception and Psychophysics*, *65*(4), 536–552. https://doi.org/10.3758/BF03194581

Lantolf, J. P. (2006). Re(de)fining language proficiency in light of the concept of 'languaculture.' In H. Byrnes (Ed.), *Advanced language learning: The contribution of Halliday and Vygotsky* (pp. 72–91). Bloomsbury.

Lantolf, J. P., Thorne, S. L., & Poehner, M. E. (2020). Sociocultural theory and L2 development. In B. VanPatten, G. D. Keating & S. Wulff (Eds.), *Theories in second language acquisition* (pp. 223–247). Routledge. https://doi.org/10.4324/9780429503986-10

Larson, J. W. (1984). Testing speaking ability in the classroom: The semi-direct alternative. *Foreign Language Annals*, *17*(5), 499–507. https://doi.org/10.1111/j.1944- 9720.1984.tb01738.x

Laurent, A., & Nicoladis, E. (2015). Gesture restriction affects French–English bilinguals' speech only in French. *Bilingualism: Language and cognition*, *18*(2), 340–349. https://doi.org/10.1017/S1366728914000042

Lavolette, E. (2013). Effects of technology modes on ratings of learner recordings. *The IALLT Journal*, *43*(2), 1–27. https://doi.org/10.17161/iallt.v43i2.8524

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing*, *13*(2), 151–172. https://doi.org/10.1177/026553229601300202

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Levelt, W. J. (1993). *Speaking: From intention to articulation*. https://doi.org/10.7551/mitpress/6393.001.0001

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–38. https://doi.org/10.1017/S0140525X99001776

Levenson, R. W., & Ekman, P. (2002). Difficulty does not account for emotion-specific heart rate changes in the directed facial action task. *Psychophysiology*, *39*(3), 397–405. https://doi.org/10.1017/S0048577201393150

Levenson, R. W., Ekman, P., & Friesen, W. V. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*, *27*(4), 363–384. https://doi.org/10.1111/j.1469-8986.1990.tb02330.x

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. *Trends in Cognitive Sciences*, *20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Levy, R. (1973). *Tahitians*. Chicago University Press.

Li, C., Dewaele, J.-M., & Jiang, G. (2020). The complex relationship between classroom emotions and EFL achievement in China. *Applied Linguistics Review*, *11*(3), 485–510. http://doi.org/10.1515/applirev-2018-0043

Li, H. Z. (2006). Backchannel responses as misleading feedback in intercultural discourse. *Journal of Intercultural Communication Research*, *35*(2), 99–116. https://doi.org/10.1080/17475750600909253

Lim, G. (2011). The development and maintenance of rating quality in performance writing assessments: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543–560. https://doi.org/10.1177/0265532211406422

Lim, G. S., Geranpayeh, A., Khalifa, H., & Buckendahl, C. W. (2013). Standard setting to an international reference framework: Implications for theory and practice. *International Journal of Testing*, *13*(1), 32–49. https://doi.org/10.1080/15305058.2012.678526

Lin, Y. (2022). Speech-accompanying gestures in L1 and L2 conversational interaction by speakers of different proficiency levels. *International Review of Applied Linguistics in Language Teaching*, *60*(2), 123–142. https://doi.org/10.1515/iral-2017-0043

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878. https://www.rasch.org/rmt/rmt162.pdf

Linacre, J. M. (2021). *Facets Rasch measurement computer program* (Version 3.83.6) [Computer software]. https://www.winsteps.com/facets.htm

Linacre, M. (n.d.). *Inter-rater and intra-rater reliability.* https://www.winsteps.com/facetman/inter-rater-reliability.htm

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. (2022). Second language anxiety in conversation and its relationship with speakers' perceptions of the interaction and their social networks. *Studies in Second Language Acquisition*. Advanced online publication. https://doi.org/10.1017/S0272263122000523

Lischetzke, T., & Eid, M. (2003). Is attention to feelings beneficial or detrimental to affective well-being? Mood regulation as a moderator variable. *Emotion*, *3*(4), 361–377. https://doi.org/10.1037/1528-3542.3.4.361

Lochner, K. (2016). *Successful emotions: How emotions drive cognitive performance*. Springer. https://doi.org/10.1007/978-3-658-12231-7

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.

Luk, J. (2010). Talking to score: Impression management in L2 oral assessment and the co-construction of a test discourse genre. *Language Assessment Quarterly, 7*(1), 25–53. https://doi.org/10.1080/15434300903473997

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing, 19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa

MacIntyre, P. D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning, 44*(2), 283–305. https://doi.org/10.1111/j.1467-1770.1994.tb01103.x

MacIntyre, P. D., & Gregersen, T. (2012). Emotions that facilitate language learning: The positive-broadening power of the imagination. *Studies in Second Language Learning and Teaching, 2*(2), 193–213. http://doi.org/10.14746/ssllt.2012.2.2.4

MacIntyre, P. D., Gregersen, T., & Mercer, S. (2019). Setting an agenda for positive psychology in SLA: Theory, practice, and research. *The Modern Language Journal, 103*(1), 262–274. https://doi.org/10.1111/modl.12544

MacIntyre, P., Clément, R., Dörnyei, Z., & Noels, K. (1998). Conceptualising willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *Modern Language Journal, 82*(4), 545–562. https://doi.org/10.1111/j.1540-4781.1998.tb05543.x

MacIntyre, P., Noels, K., & Clément, R. (1997). Biases in self- ratings of second language proficiency: The role of language anxiety. *Language Learning, 47*(2), 265–287. https://doi.org/10.1111/0023-8333.81997008

Manstead, A. S., & Wagner, H. L. (1981). Arousal, cognition and emotion: An appraisal of two-factor theory. *Current Psychological Reviews, 1*(1), 35–54. https://doi.org/10.1007/BF02979253

Marefat, F., & Heydari, M. (2016). Native and Iranian teachers' perceptions and evaluation of Iranian students' English essays. *Assessing Writing, 27*, 24–36. https://doi.org/10.1016/j.asw.2015.10.001

Mark, G., Knight, D., O'Keeffe, A., & Fitzgerald, C. (2023). *Interactional Variation Online (IVO): Corpus approaches and applications to analyzing multi-modal collaboration in virtual meetings* [Paper presentation]. American Association of Applied Linguistics. Portland, Oregon, USA.

Marsh, A. A., Elfenbein, H. A., & Ambady, N. (2003). Nonverbal "accents" cultural differences in facial expressions of emotion. *Psychological Science, 14*(4), 373–376. https://doi.org/10.1111/1467-9280.24461

Masuda, T., Ellsworth, P. C., Mesquita, B., Leu, J., Tanida, S., & Van de Veerdonk, E. (2008). Placing the face in context: Cultural differences in the perception of facial emotion. *Journal of Personality and Social Psychology, 94*(3), 365–381. https://doi.org/10.1037/0022-3514.94.3.365

Masuda, T., Wang, H., Ishii, K., & Ito, K. (2012). Do surrounding figures' emotions affect judgment of the target figure's emotion? Comparing the eye-movement patterns of European Canadians, Asian Canadians, Asian international students, and Japanese. *Frontiers in Integrative Neuroscience, 6*, Article 72. https://doi.org/10.3389/fnint.2012.00072

Matsumoto, D. (2001). Culture and emotion. In D. Matsumoto (Ed.), *The handbook of culture and psychology* (pp. 171–194). Oxford University Press.

Matsumoto, D., & Hwang, H. C. (2016). The cultural bases of nonverbal communication. In D. Matsumoto, H. C. Hwang, & M. G. Frank (Eds.), *APA handbook of nonverbal communication* (pp. 77–101). American Psychological Association. https://doi.org/10.1037/14669-004

Matsumoto, D., & Hwang, H. S. (2012). Culture and emotion: The integration of biological and cultural contributions. *Journal of Cross-Cultural Psychology*, *43*(1), 91–118. https://doi.org/10.1177/0022022111420147

Matsumoto, D., Hwang, H. C., & Frank, M. G. (Eds.) (2016). *APA handbook of nonverbal communication*. American Psychological Association. https://doi.org/10.1037/14669-000

Matsumoto, D., Olide, A., Schug, J., Willingham, B., & Callan, M. (2009). Cross-cultural judgments of spontaneous facial expressions of emotion. *Journal of Nonverbal Behavior*, *33*, 213–238. https://doi.org/10.1007/s10919-009-0071-4

Matsumoto, Y. (2018). Functions of laughter in English-as-a-lingua-franca classroom interactions: A multimodal ensemble of verbal and nonverbal interactional resources at miscommunication moments. *Journal of English as a Lingua Franca*, *7*(2), 229–260. https://doi.org/10.1515/jelf-2018-0013

Max Planck Institute. (2020). *ELAN* (Version 5.9) [Computer software]. The Language Archive.

May, L. (2009). Co-constructed interaction in a paired speaking test: The raters' perspective. *Language Testing*, *26*(3), 397–421. https://doi.org/10.1177/0265532209104668

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127–145. https://doi.org/10.1080/15434303.2011.565845

McCafferty, S. G. (2006). Gesture and the materialization of second language prosody. *International Review of Applied Linguistics*, *44*(2), 195–207. https://doi.org/10.1515/IRAL.2006.008

McCafferty, S. G., & Ahmed, M. K. (2000). The appropriation of gestures of the abstract by L2 learners. In J. P. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 199–218). Oxford University Press.

McDonough, K., Crowther, D., Kielstra, P., & Trofimovich, P. (2015). Exploring the potential relationship between eye gaze and English L2 speakers' responses to recasts. *Second Language Research*, *31*(4), 563–575. https://doi.org/10.1177/0267658315589656

McDonough, K., Kim, Y. L., Uludag, P., Liu, C., & Trofimovich, P. (2022a). Exploring the relationship between behavior matching and interlocutor perceptions in L2 interaction. *System*, *109*, Article 102865. https://doi.org/10.1016/j.system.2022.102865

McDonough, K., Lindberg, R., & Trofimovich, P. (2022b). Examining rater perception of holds as a visual cue of listener nonunderstanding. *Studies in Second Language Acquisition*, *44*(5), 1240–1259. https://doi.org/10.1017/S0272263122000018

McDonough, K., Lindberg, R., Trofimovich, P., & Tekin, O. (2023). The visual signature of non-understanding: A systematic replication of McDonough, Trofimovich, Lu, and Abashidze (2019). *Language Teaching*, *56*(1), 113–127. https://doi.org/10.1017/S0261444821000197

McDonough, K., Trofimovich, P., Lu, L., & Abashidze, D. (2019). The occurrence and perception of listener visual cues during nonunderstanding episodes. *Studies in Second Language Acquisition*, *41*(5), 1151–1165. https://doi.org/10.1017/S0272263119000238

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

McNamara, T. (1996). *Measuring second language performance*. Longman.

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice and language assessment*. Oxford.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, *92*(3), 350–371. https://doi.org/10.1037/0033-295X.92.3.350

McNeill, D. (1992) *Hand and mind: What the hands reveal about thought*. University of Chicago Press.

McNeill, D. (2005). *Gesture & thought*. The University of Chicago Press. https://doi.org/10.7208/chicago/9780226514642.001.0001

McNeill, D., & Duncan, S. (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 141–161). Cambridge University Press. https://doi.org/10.1017/CBO9780511620850.010

Mehrabian, A. (1972). *Nonverbal communication.* Aldine-Atherton.

Mehrabian, A. (1981). *Silent messages: Implicit communication of emotions and attitudes* (2nd ed.). Wadsworth.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, *14*, 261–292. https://doi.org/10.1007/BF02686918

Mehrabian, A., & Williams, M. (1969). Nonverbal concomitants of perceived and intended persuasiveness. *Journal of Personality and Social Psychology, 13*(1), 37–58. https://doi.org/10.1037/h0027993

Meltzoff, A. N., & Moore, M. K. (1997). Explaining facial imitation: A theoretical model. *Infant and Child Development*, *6*(3-4), 179–192. https://doi.org/10.1002/(SICI)1099-0917(199709/12)6:3/4%3C179::AID-EDP157%3E3.0.CO;2-R

Mesquita, B. (2022). *Between us: How cultures create emotions*. W. W. Norton & Company.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Mondada, L. (2006). Participants' online analysis and multimodal practices: projecting the end of the turn and the closing of the sequence. *Discourse Studies*, *8*(1), 117–129. https://doi.org/10.1177/1461445606059561

Mondada, L. (2007). Multimodal resources for turn-taking: Pointing and the emergence of possible next speakers. *Discourse Studies*, *9*(2), 194–225. https://doi.org/10.1177/1461445607075346

Mondada, L. (2014). The local constitution of multimodal resources for social interaction. *Journal of Pragmatics*, *65*, 137–156. https://doi.org/10.1016/j.pragma.2014.04.004

Mondada, L. (2016). Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics*, *20*(3), 336–366. https://doi.org/10.1016/j.pragma.2014.04.004

Morett, L. M. (2014). When hands speak louder than words: The role of gesture in the communication, encoding, and recall of words in a novel second language. *The Modern Language Journal*, *98*(3), 834–853. https://doi.org/10.1111/modl.12125

Morgenstern, A., & Goldin-Meadow, S. (Eds.). (2022). Introduction to gesture in language. In A. Morgenstern & S. Goldin-Meadow (Eds.), *Gesture in language: Development across the lifespan* (pp. 3–17). De Gruyter Mouton. https://doi.org/10.1037/0000269-001

Morreale, S. P., Spitzbert, B. H., & Barge, J. K. (2013). *Communication: motivation, knowledge, skills* (3rd Ed.). Peter Lang. https://doi.org/10.3726/978-1-4539-0257-8

Morris, D., Collet, P., Marsh, P., & O'Shaughnessy, M. (1979). *Gestures: Their origin and distribution*. Stein and Day Publisher.

Morris, J. S., Frith, C. D., Perrett, D. I., Rowland, D., Young, A. W., Calder, A. J., & Dolan, R. J. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature*, *383*(6603), 812–815. https://doi.org/10.1038/383812a0

Morrow, K. (1979). Communicative language testing: Revolution or evolution? In C. J. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 143–159). Oxford University Press.

Morsbach, H., & Tyler, W. J. (1986). A Japanese emotion: Amae. In R. Harré (Ed.), *The social construction of emotion* (pp. 289–308). Blackwell.

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*(3), 289–306. https://doi.org/10.1177/002383099503800305

Nagle, C. L., Trofimovich, P., O'Brien, M. G., & Kennedy, S. (2022). Beyond linguistic features: Exploring behavioral and affective correlates of comprehensible second language speech. *Studies in Second Language Acquisition*, *44*(1), 255–270. https://doi.org/10.1017/S0272263121000073

Nakatsuhara, F. (2011). Effects of test taker characteristics and the number of participants in group oral tests. *Language Testing*, *28*(4), 483–508. https://doi.org/10.1177/0265532211398110

Nakatsuhara, F., Inoue, C., & Taylor, L. (2021a). Comparing rating modes: Analyzing live, audio, and video ratings of IELTS speaking test performances. *Language Assessment Quarterly*, *18*(2), 83–106. https://doi.org/ 10.1080/15434303.2020.1799222

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly*, *14*(1), 1–18. https://doi.org/10.1080/15434303.2016.1263637

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2021b). Video-conferencing speaking tests: Do they measure the same construct as face-to-face tests? *Assessment in Education: Principles, Policy & Practice*, *28*(4), 369–388. https://doi.org/10.1080/0969594X.2021.1951163

Nakatsukasa, K. (2016). Efficacy of recasts and gestures on the acquisition of locative prepositions. *Studies in Second Language Acquisition*, *38*(4), 771–799. https://doi.org/10.1017/S0272263115000467

Nambiar, M. K., & Goon, C. (1993). Assessment of oral skills : A comparison of scores obtained through audio recordings to those obtained through face-to-face evaluation. *RELC Journal*, *24*(1), 15–31. https://doi.org/10.1177/003368829302400102

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin*, *35*(12), 1661–1671. https://doi.org/10.1177/0146167209346309

Negueruela, E., Lantolf, J. P., Jordan, S. R., & Gelabert, J. (2004). The "private function" of gesture in second language speaking activity: a study of motion verbs and gesturing in English and Spanish. *International Journal of Applied Linguistics*, *14*(1), 113-147. https://doi.org/10.1111/j.1473-4192.2004.00056.x

Nestler, S., & Back, M. D. (2013). Applications and extensions of the lens model to understand interpersonal judgments at zero acquaintance. *Current Directions in Psychological Science*, *22*(5), 374–379. https://doi.org/10.1177/0963721413486148

Neu, J. (1990). Assessing the role of nonverbal communication in the acquisition of communicative competence in L2. In R. C. Scarcella, E. S. Andersen, & S. D. Krashen (Eds.), *Developing communicative competence in a second language* (pp. 121–138). Newbury House.

Nicoladis, E. (2007). The effect of bilingualism on the use of manual gestures. *Applied Psycholinguistics*, *28*(3), 441-454. https://doi.org/10.1017/S0142716407070245

Nicoladis, E. (2007). The effect of bilingualism on the use of manual gestures. *Applied Psycholinguistics*, *28*(3), 441–454. https://doi.org/10.1017/S0142716407070245

Nicoladis, E., Mayberry, R. I., & Genesee, F. (1999). Gesture and early bilingual development. *Developmental Psychology*, *35*(2), 514–526. https://doi.org/10.1037/0012-1649.35.2.514

Nicoladis, E., Pika, S., Yin, H. U., & Marentette, P. (2007). Gesture use in story recall by Chinese–English bilinguals. *Applied Psycholinguistics*, *28*(4), 721–735. https://doi.org/10.1017/S0142716407070385

Noels, K. A., & Clément, R. (1996). Communicating across cultures: Social determinants and acculturative consequences. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *28*(3), 214–228. https://doi.org/10.1037/0008-400X.28.3.214

Noels, K. A., Pon, G., & Clément, R. (1996). Language, identity, and adjustment: The role of linguistic self-confidence in the acculturation process. *Journal of Language and Social Psychology*, *15*(3), 246–264. https://doi.org/10.1177/0261927X960153003

Norton, B. (2000). *Identity and language learning: Gender, ethnicity and educational change*. Pearson.

Norton, B. (2013). *Identity and language learning: Extending the conversation*. Multilingual Matters. https://doi.org/10.21832/9781783090563

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.

O'Sullivan, B. (1996). The evaluation of gestures in non-verbal communication. In G. van Troyer, S. Cornwell, & H. Morikawa (Eds.), *Proceedings of the JALT 1995 conference* (pp. 316–320). The Japan Association for Language Teaching. https://files.eric.ed.gov/fulltext/ED402769.pdf

O'Sullivan, B. (2004). Modelling factors affecting oral language test performance: A large scale empirical study. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context: Studies in Language Testing 18* (pp. 129–142). Cambridge University Press.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24*(4), 517–537. https://doi.org/10.1177/0265532207080771

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing*, *26*(2), 161–186. https://doi.org/10.1177/0265532208101005

Oloff, F. (2018). "Sorry?"/"Como?"/"Was?" – Open class and embodied repair initiators in international workplace interactions. *Journal of Pragmatics*, *126*, 29–51. https://doi.org/10.1016/j.pragma.2017.11.002

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*(2), 143–154. https://doi.org/10.1016/S0346-251X(02)00002-7

Osgood, C. E., Suci, G. J., & Tannenbaum, P. (1957). *The measurement of meaning*. University of Illinois Press.

Otero, S. C., Weekes, B. S., & Hutton, S. B. (2011). Pupil size changes during recognition memory. *Psychophysiology*, *48*(10), 1346–1353. https://doi.org/10.1111/j.1469-8986.2011.01217.x

Oxford, R. L. (2016). Toward a psychology of well-being for language learners: The "EMPATHICS" vision. In T. Gregersen, M. D. MacIntyre, & S. Mercer (Eds.), *Positive psychology in SLA* (pp. 10–87). Multilingual Matters. https://doi.org/10.21832/9781783095360-003

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society B*, *369*(1651), Article 20130296. https://doi.org/10.1098/rstb.2013.0296

Özyürek, A., & Kita, S. (1999). Expressing manner and path in English and Turkish: Differences in speech, gesture, and conceptualization. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the 21st cognitive science meeting* (pp. 507–512). Lawrence Erlbaum Associates, Publishers. https://doi.org/10.4324/9781410603494-94

Özyürek, A., & Kelly, S. D. (2007). Gesture, language, and brain. *Brain and Language*, *101*(3), 181–185. https://doi.org/10.1016/j.bandl.2007.03.006

Pallotti, G. (2021). Measuring complexity, accuracy, and fluency (CAF). In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 201–210). Routledge. https://doi.org/10.4324/9781351034784-23

Pan, M. (2016). *Nonverbal delivery in speaking assessment: From an argument to a rating scale formulation and validation*. Springer. https://doi.org/10.1007/978-981-10-0170-3

Parkinson, B. (1996). Emotions are social. *British journal of Psychology*, *87*(4), 663–683. https://doi.org/10.1111/j.2044-8295.1996.tb02615.x

Parkinson, B. (2011). Interpersonal emotion transfer: Contagion and social appraisal. *Social and Personality Psychology Compass*, *5*(7), 428-439. https://doi.org/10.1111/j.1751-9004.2011.00365.x

Parkinson, B. (2019). *Heart to heart: How your emotions affect other people*. Cambridge University Press. https://doi.org/10.1017/9781108696234

Parkinson, B., & Simons, G. (2009). Affecting others: Social appraisal and emotion contagion in everyday decision making. *Personality and Social Psychology Bulletin*, *35*(8), 1071–1084. https://doi.org/ 10.1177/0146167209336611

Patterson, M. L. (1973). Compensation in nonverbal immediacy behaviors: A review. *Sociometry, 36*(2), 237–252. https://doi.org/10.2307/2786569

Patterson, M. L. (1983). *Nonverbal Behavior: A Functional Perspective*. Springer-Verlag. https://doi.org/10.1007/978-1-4612-5564-2

Patterson, M. L., & Ritts, V. (1997). Social and communicative anxiety: A review and meta-analysis. *Annals of the International Communication Association*, *20*(1), 263–303. https://doi.org/10.1080/23808985.1997.11678944

Pavlenko, A. (2006). *Bilingual minds: Emotional experience, expression, and representation* (Vol. 56). Multilingual Matters. https://doi.org/10.21832/9781853598746

Pavlenko, A. (2014). *The bilingual mind: And what it tells us about language and thought*. Cambridge University Press. https://doi.org/10.1017/CBO9781139021456

Pavlenko, A., & Norton, B. (2007). Imagined communities, identity, and English language learning. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 669-680). Springer. https://doi.org/10.1007/978-0-387-46301-8_43

Pekarek Doehler, S., & Berger, E. (2018). L2 interactional competence as increased ability for context-sensitive conduct: A longitudinal study of story-openings. *Applied Linguistics, 39*(4), 555–578. https://doi.org/10.1093/applin/amw021

Pekarek Doehler, S., & Pochon-Berger, E. (2015). The development of L2 interactional competence: evidence from turn-taking organization, sequence organization, repair organization and preference organization. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-Based Perspectives on Second Language Learning* (pp. 233–268). De Gruyter Mouton. https://doi.org/10.1515/9783110378528-012

Pekarek Doehler, S., & Skogmyr Marian, K. (2022). Functional diversification and progressive routinization of a multiword expression in and for social interaction: A longitudinal L2 study. *The Modern Language Journal, 106*(S1), 23–45. https://doi.org/10.1111/modl.12758

Pennycook, A. (1985). Actions speak louder than words: Paralanguage, communication, and education. *TESOL Quarterly, 19*(2), 259–282. https://doi.org/10.2307/3586829

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

Philp, J., & Duchesne, S. (2016). Exploring engagement in tasks in the language classroom. *Annual Review of Applied Linguistics, 36,* 50–72. https://doi.org/10.1017/S0267190515000094

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169–226. https://doi.org/10.1017/S0140525X04000056

Pika, S., Nicoladis, E., & Marentette, P. (2009). How to order a beer: Cultural differences in the use of conventional gestures for numbers. *Journal of Cross-Cultural Psychology, 40*(1), 70–80. https://doi.org/10.1177/0022022108326197

Pine, K. J., Gurney, D. J., & Fletcher, B. (2010). The semantic specificity hypothesis: When gestures do not depend upon the presence of a listener. *Journal of Nonverbal Behavior, 34*, 169–178. https://doi.org/10.1007/s10919-010-0089-7

Pitzl, M. L. (2010). *English as a Lingua Franca in international business: Resolving miscommunication and reaching shared understanding.* VDM-Verlag Müller.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Plough, I. (2021). A case for nonverbal behavior: Implications for construct, performance, and assessment. In Salaberry, M. R. & Burch, A. R. (Eds.), *Assessing speaking in context—Expanding the construct and its applications* (pp. 50–72). Multilingual Matters. https://doi.org/10.21832/9781788923828-004

Plough, I. C., & Bogart, P. S. (2008). Perceptions of examiner behavior modulate power relations in oral performance testing. *Language Assessment Quarterly*, *5*(3), 195–217. https://doi.org/10.1080/15434300802229375

Plough, I., Banerjee, J., & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing*, *35*(3), 427–455. https://doi.org/10.1177/0265532218772325

Plusquellec, P., & Denault, V. (2018). The 1000 most cited papers on visible nonverbal behavior: A bibliometric analysis. *Journal of Nonverbal Behavior*, *42*(3), 347–377. https://doi.org/10.1007/s10919-018-0280-9

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist, 89(4),* 344–350. https://doi.org/10.1511/2001.28.344

Prior, M. T. (2019). Elephants in the room: An "affective turn," or just feeling our way? *The Modern Language Journal*, *103*(2), 516–527. http://doi.org/10.1111/modl.12573

Qiu, X., & Lo, Y. Y. (2017). Content familiarity, task repetition and Chinese EFL learners' engagement in second language use. *Language Teaching Research*, *21*(6), 681–698. https://doi.org/10.1177/1362168816684368

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Randez, R. A., & Cornell, C. (in press). Advancing equity in language assessment for learners with disabilities. *Language Testing*.

Rasmussen, G. (2014). Inclined to better understanding—the coordination of talk and 'leaning forward' in doing repair. *Journal of Pragmatics, 65,* 30–45. https://doi.org/10.1016/j.pragma.2013.10.001

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, *7*(4), 226–231. https://doi.org/10.1111/j.1467-9280.1996.tb00364.x

Reynolds Jr, D. A. J., & Gifford, R. (2001). The sounds and sights of intelligence: A lens model channel analysis. *Personality and Social Psychology Bulletin*, *27*(2), 187–200. https://doi.org/10.1177/0146167201272005

Richmond, V., & McCroskey, J. (2004). *Nonverbal behavior in interpersonal relationships*. Allyn and Bacon.

Rimé, B., Mesquita, B., Boca, S., & Philippot, P. (1991). Beyond the emotional event: Six studies on the social sharing of emotion. *Cognition & Emotion*, *5*(5-6), 435-465. https://doi.org/10.1080/02699939108411052

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27,* 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Roever, C. (2021). *Teaching and testing second language pragmatics and interaction.* Routledge. https://doi.org/10.4324/9780429260766

Roever, C., & Dai, D. W. (2021). Reconceptualizing interactional competence for language testing. In Salaberry, M. R. & Burch, A. R. (Eds.), *Assessing speaking in context—Expanding the construct and its applications* (pp. 23–40). Multilingual Matters. https://doi.org/10.21832/9781788923828-003

Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing*, *35*(3), 331–355. https://doi.org/10.1177/0265532218758128

Ross, S. (1992). Accommodative questions in oral proficiency interviews. *Language Testing*, *9*(2), 173–185. https://doi.org/10.1177/026553229200900205

Rossano, F. (2012). *Gaze behavior in face-to-face interaction* [Unpublished doctoral dissertation]. Radboud University. http://hdl.handle.net/2066/99151

Rost, M. (2016). *Teaching and researching listening* (3rd ed.). Taylor and Francis. https://doi.org/10.4324/9781315732862

Rule, N. O., & Alaei, R. (2016). "Gaydar" the perception of sexual orientation from subtle cues. *Current Directions in Psychological Science*, *25*(6), 444–448. https://doi.org/10.1177/0963721416664403

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, *115*(1), 102–141. https://doi.org/10.1037/0033-2909.115.1.102

Russell, J. A. (2012). Introduction to special section: On defining emotion. *Emotion Review*, *4*(4), 337–337. https://doi.org/10.1177/1754073912445857

Russell, J. A., & Fehr, B. (1987). Relativity in the perception of emotion in facial expressions. *Journal of Experimental Psychology: General*, *116*(3), 223–237. https://doi.org/10.1037/0096-3445.116.3.223

Rylander, J., Clark, P., & Derrah, R. (2013). A video-based method of assessing pragmatic awareness. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 65–97). Springer. https://doi.org/10.1057/9781137003522_3

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735. https://doi.org/10.2307/412243

Saito, K., Ilkan, M., Magne, V., Tran, M. N., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid-and high-level second language fluency. *Applied Psycholinguistics*, *39*(3), 593–617. https://doi.org/10.1017/S0142716417000571

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition*, *19*(3), 597–609. https://doi.org/10.1017/S1366728915000255

Sato, T., & McNamara, T. (2019). What counts in second language oral communication ability? The perspective of linguistic laypersons. *Applied Linguistics*, *40*(6), 894–916. https://doi.org/10.1093/applin/amy032

Sauter, D. A., & Eimer, M. (2010). Rapid detection of emotion from human vocalizations. *Journal of Cognitive Neuroscience*, *22*(3), 474–481. https://doi.org/10.1162/jocn.2009.21215

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, *107*(6), 2408–2412. https://doi.org/10.1073/pnas.0908239106

Savignon, S. J. (1972). *Communicative competence: An experiment in foreign-language teaching.* Center for Curriculum Development.

Scarcella, R. C., Andersen, E. S., & Krashen, S. D. (1990). *Developing communicative competence in a second language*. Newbury House Publishers.

Schachter, S. (1964). The interaction of cognitive and physiological determinants of emotional state. *Advances in Experimental Social Psychology*, *1*, 49–80. https://doi.org/10.1016/S0065-2601(08)60048-9

Schegloff, E. A. (2006) Interaction: The infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of human society* (pp. 70–96). Berg. https://doi.org/10.4324/9781003135517-4

Scherer, K. R. (2005). What are emotions? *Social Science Information*, *44*(4), 695–729. https://doi.org/10.1177/0539018405058216

Schieffelin, B. B., & Ochs, E. (Eds.) (1979). *Developmental pragmatics*. Academic Press.

Schmid Mast, M., & Cousin, G. (2013). Power, dominance, and persuasion. In J. A. Hall & M. L. Knapp (Eds.), *Nonverbal communication* (pp. 613–635). De Gruyter. https://doi.org/10.1515/9783110238150.613

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing*, *22*(1), 1–30. https://doi.org/10.1191/0265532205lt295oa

Scovel, T. (1978). The effect of affect on foreign language learning: A review of the anxiety research. *Language Learning*, *28*(1), 129–142. https://doi.org/10.1111/j.1467-1770.1978.tb00309.x

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge. https://doi.org/10.4324/9780203851357

Seligman, M. E. P. (2011). *Flourish: A visionary new understanding of happiness and well-being.* Free Press.

Şen, M., & Oz, H. (2021). Vocabulary size as a predictor of willingness to communicate inside the classroom. In N. Zarrinabadi & M. Pawlak (Eds.), *New perspectives on willingness to communicate in a second language* (pp. 235–259). Springer. https://doi.org/10.1007/978-3-030-67634-6_12

Seo, M. S., & Koshik, I. (2010). A conversation analytic study of gestures that engender repair in ESL conversational tutoring. *Journal of Pragmatics*, *42*(8), 2219–2239. https://doi.org/10.1016/j.pragma.2010.01.021

Sevinç, Y. (2018). Language anxiety in the immigrant context: Sweaty palms? *International Journal of Bilingualism*, *22*(6), 717–739. https://doi.org/10.1177/1367006917690914

Seyfeddinipur, M. (2006). *Disfluency: Interrupting speech and gesture* [Unpublished doctoral dissertation]. Radboud University.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, *52*(6), 1061–1086. https://doi.org/10.1037/0022-3514.52.6.1061

Shaw, S. (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Cambridge Research Notes*, *8*(5), 13–17. https://www.cambridgeenglish.org/Images/23120-research-notes-08.pdf

Sherman, J., & Nicoladis, E. (2004). Gestures by advanced Spanish–English second-language learners. *Gesture*, *4*(2), 143–156. https://doi.org/10.1075/gest.4.2.03she

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluations of Chinese students' English writing. *Language Testing*, *18*(3), 303–325. https://doi.org/10.1177/026553220101800303

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*(4), 489–503. https://doi.org/10.1515/ppb-2017-0056

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, *11*(2), 99–123. https://doi.org/10.1177/026553229401100202

Siegel, E. H., Sands, M. K., Van den Noortgate, W., Condon, P., Chang, Y., Dy, J., Quigley, K. S., & Barrett, L. F. (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychological Bulletin*, *144*(4), 343–393. https://doi.org/10.1037/bul0000128

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*(4), 557–566. https://doi.org/10.1037/pas0000648

Singelis, T. (1994). Nonverbal communication in intercultural interactions. In R. Brislin & T. Yoshida (Eds.), *Improving intercultural interactions* (pp. 268–294). Sage. https://doi.org/10.4135/9781452204857.n14

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press. https://doi.org/10.1177/003368829802900209

Skogmyr Marian, K. (2023). *The development of L2 interactional competence: A multimodal study of complaining in French interactions.* Routledge. https://doi.org/10.4324/9781003271215

Slobin, D. I. (1996). Two ways to travel: Verbs of motion in English and Spanish. In M. Shibatani & S. A. Thompson (Eds.), *Grammatical constructions: Their form and meaning* (pp. 195–220). Clarendon Press.

Slobin, D. I. (2006). What makes manner of motion salient? Explorations in linguistic typology, discourse, and cognition. In M. Hickmann & S. Robert (Eds.), *Space in languages: Linguistic systems and cognitive categories* (pp. 59–81). https://doi.org/10.1075/tsl.66.05slo

Smith, C. A., & Lazarus, R. S. (1993). Appraisal components, core relational themes, and the emotions. *Cognition & Emotion*, *7*(3-4), 233–269. https://doi.org/10.1080/02699939308409189

Snider, J. G., & Osgood, C. E. (Eds.). (1969). *Semantic differential technique: A sourcebook*. Aldine.

So, W. C., Kita, S., & Goldin-Meadow, S. (2013). When do speakers use gestures to specify who does what to whom? The role of language proficiency and type of gestures in narratives. *Journal of Psycholinguistic Research*, *42*, 581–594. https://doi.org/10.1007/s10936-012-9230-6

Spielberger, C. D. (1983). *Manual for the State-Trait-Anxiety: STAI (form Y)*. Consulting Psychologists Press.

Stam, G. (1998). Changes in patterns of thinking about motion with L2 acquisition. In S. Santi, I. Guaïtella, C. Cavé, & G. Konopczynski (Eds.), *Oralité et gestualité: Communication multimodale, interaction* (pp. 615–619). L'Harmattan.

Stam, G. (2006). Thinking for speaking about motion: L1 and L2 speech and gesture. *IRAL*, *44*(2), 145–171. https://doi.org/10.1515/IRAL.2006.006

Stam, G. (2008). What gestures reveal about second language acquisition. In S. G. McCafferty & S. Stam (Eds.). *Gesture: Second language acquisition and classroom research* (pp. 231–256). Routledge.

Stam, G. (2010). *Can a L2 speaker's patterns of thinking for speaking change?* In Z. H. Han & T. Cadierno (Eds.), Linguistic relativity in SLA: Thinking for speaking (pp. 59–83). Multilingual Matters. https://doi.org/10.21832/9781847692788-005

Stam, G. (2017). Verb-framed, satellite framed, or in between? A L2 learner's thinking for speaking in her L1 and L2 over 14 years. In I. Ibarretxe-Antuñano (Ed.), *Motion and space across languages: Theory and applications* (pp. 329–366). John Benjamins. https://doi.org/10.1075/hcp.59.14sta

Stam, G., & Tellier, M. (2022). Gesture helps second and foreign language learning and teaching. In A. Morgenstern & S. Goldin-Meadow (Eds.), *Gesture in language: Development across the lifespan* (pp. 335–363). American Psychological Association. https://doi.org/10.1037/0000269-014

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety?. *Learning and Individual Differences*, *22*(6), 747–758. https://doi.org/10.1016/j.lindif.2012.05.013

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592. https://doi.org/10.1073/pnas.090361610

Stöckli, S., Schulte-Mecklenbeck, M., Borer, S., & Samson, A. C. (2018). Facial expression analysis with AFFDEX and FACET: A validation study. *Behavior Research Methods*, *50*, 1446–1460. https://doi.org/10.3758/s13428-017-0996-1

Streeck, J. (2009). Forward-gesturing. *Discourse Processes*, *46*(2–3), 161–179. https://doi.org/10.1080/01638530902728793

Streeck, J., & Hartge, U. (1992). Previews: Gestures at the transition place. In P. Auer & A. di Luzio (Eds.), *The contextualization of language* (pp. 135–157). Benjamins. https://doi.org/10.1075/pbns.22.10str

Styles, P. (1993). *Inter- and intra-rater reliability of assessments of "live" versus audio- and video-recorded interviews in the IELTS Speaking test*. British Council.

Sueyoshi, A., & Hardison, D. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661–669. https://doi.org/10.1111/j.0023-8333.2005.00320.x

Suslow, T., P. Ohrmann, J. Bauer, A. V. Rauch, W. Schwindt, V. Arolt, W. Heindel, & H. Kugel. (2006). Amygdala activation during masked presentation of emotional faces predicts conscious detection of threat-related faces. *Brain and Cognition*, *61*(3), 243–248. https://doi.org/10.1016/j.bandc.2006.01.005

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing*, *32*(4), 463–483. https://doi.org/10.1177/0265532214562099

Suvorov, R. (2018). Test takers' use of visual information in an L2 video-mediated listening test: Evidence from cued retrospective reporting. In E. Wagner & G. J. Ockey (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 145–160). John Benjamins. https://doi.org/10.1075/lllt.50.10suv

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143–167. https://doi.org/10.1017/S0272263119000421

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal*, *105*(2), 435–463. https://doi.org/10.1111/modl.12706

Tabachnik, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description: Vol. 3. Grammatical categories and the lexicon* (pp. 57–149). Cambridge University Press.

Talmy, L. (2000). *Towards a cognitive semantics: Vol. II: Typology and process in concept structuring*. MIT Press. https://doi.org/10.7551/mitpress/6848.001.0001

Tavakoli, P., & Skehan, P. (2005). 9. Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-273). John Benjamins. https://doi.org/10.1075/lllt.11.15tav

Taylor, L. B., & Banerjee, J. (in press). Accommodations in language testing and assessment: Safeguarding equity, access, and inclusion. *Language Testing*.

Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, *41*(2), 363–387. https://doi.org/10.1017/S0272263118000311

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, *8*(2), 219–235. https://doi.org/10.1075/gest.8.2.06tel

The National Standards Collaborative Board. (2015). *World-readiness standards for learning languages* (4th ed.). The National Standards Collaborative Board.

Thompson, C. P. (2016). *Preliminary study of the role of eye contact, gestures, and smiles produced by Chinese-as-a-first-language test takers on ratings assigned by English-as-a-first- language examiners during IELTS speaking tests* [Unpublished MA thesis]. University of Victoria, Canada. http://hdl.handle.net/1828/7724

Thompson, G. L., Cox, T. L., & Knapp, N. (2016). Comparing the OPI and the OPIc: The effect of test method on oral proficiency scores and student preference. *Foreign Language Annals*, *49*(1), 75–92. https://doi.org/10.1111/flan.12178

Tiedens, L. Z., & Fragale, A. R. (2003). Power moves: Complementarity in dominant and submissive nonverbal behavior. *Journal of Personality and Social Psychology*, *84*(3), 558–568. https://doi.org/10.1037/0022-3514.84.3.558

Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton University Press. https://doi.org/10.1515/9781400885725

Tominaga, W. (2013). The development of extended turns and storytelling in the Japanese oral proficiency interview. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 220–257). Palgrave Macmillan. https://doi.org/10.1057/9781137003522_9

Tomkins, S. S., & McCarter, R. (1964). What and where are the primary affects? Some evidence for a theory. *Perceptual and Motor Skills*, *18*(1), 119–158. https://doi.org/10.2466/pms.1964.18.1.119

Tran, V. (2007). The use, overuse, and misuse of affect, mood, and emotion in organizational research. In C. E. J. Härtel, N. M. Ashkanasy, & W. J. Zerbe (Eds.), *Functionality, intentionality and morality* (pp. 31–53). Elsevier. https://doi.org/10.1016/S1746-9791(07)03002-7

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*(4), 905–916. https://doi.org/10.1017/S1366728912000168

Trofimovich, P., Tekin, O., & McDonough, K. (2021). Task engagement and comprehensibility in interaction: Moving from what second language speakers say to what they do. *Journal of Second Language Pronunciation*, *7*(3), 435–461. https://doi.org/10.1075/jslp.21006.tro

Tsunemoto, A., Lindberg, R., Trofimovich, P., & McDonough, K. (2022). Visual cues and rater perceptions of second language comprehensibility, accentedness, and fluency. *Studies in Second Language Acquisition*, *44*(3), 659–684. https://doi.org/10.1017/S0272263121000425

Uchida, Y., & Kitayama, S. (2009). Happiness and unhappiness in east and west: Themes and variations. *Emotion, 9*(4), 441–456. https://doi.org/10.1037/a0015634

Uchida, Y., Townsend, S. S., Rose Markus, H., & Bergsieker, H. B. (2009). Emotions as within or between people? Cultural variation in lay theories of emotion expression and inference. *Personality and Social Psychology Bulletin*, *35*(11), 1427–1439. https://doi.org/10.1177/0146167209347322

Uludag, P., McDonough, K., & Trofimovich, P. (2022). Exploring shared and individual assessment of paired oral interactions. *Studies in Language Assessment, 11*(2), 1–24. studihttps://www.altaanz.org/uploads/5/9/0/8/5908292/1._sila_11_2__uludag_et_al..pdf

van Compernolle, R. A. (2013). Interactional competence and the dynamic assessment of L2 pragmatic abilities. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 327–353). Palgrave Macmillan. https://doi.org/10.1057/9781137003522_13

van der Wel, P., & Van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review, 25*, 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

Van Ek, J. A. (1986). *Objectives for foreign language learning. Volume I: Scope*. Council of Europe.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Walther, J. B., & Tidwell, L. C. (1995). Nonverbal cues in computer-mediated communication, and the effect of chronemics on relational communication. *Journal of Organizational Computing, 5*(4), 355–378. https://doi.org/10.1080/10919399509540258

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS Scales. *Journal of Personality and Social Psychology, 54*, 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Watson, O. M., & Graves, T. D. (1966). Quantitative research in proxemic behavior 1. *American Anthropologist, 68*(4), 971–985. https://doi.org/10.1525/aa.1966.68.4.02a00070

Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly*, *12*(3), 283–304. https://doi.org/10.1080/15434303.2015.1037446

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197–223. https://doi.org/10.1177/026553229401100206

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287. https://doi.org/10.1177/026553229801500205

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

WIDA. (2020). *WIDA English language development standards framework, 2020 edition: Kindergarten–grade 12*. Board of Regents of the University of Wisconsin System. https://wida.wisc.edu/sites/default/files/resource/WIDA-ELD-Standards-Framework-2020.pdf

Wierzbicka, A. (1992). Talking about emotions: Semantics, culture, and cognition. *Cognition & Emotion*, *6*(3-4), 285–319. https://doi.org/10.1080/02699939208411073

Wierzbicka, A. (1994). Emotion, language, and cultural scripts. In S. Kitayama & H. R. Markus (Eds.), *Emotion and culture: Empirical studies of mutual influence* (pp. 133–196). American Psychological Association. https://doi.org/10.1037/10152-004

Willis, J., & Todorov, A. (2006). First Impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Winke, P., Zhang, X., & Pierce, S. (2022). A closer look at a marginalized test method: Self-assessment as a measure of speaking proficiency. *Studies in Second Language Acquisition,* Advanced online publication. https://doi.org/10.1017/S0272263122000079

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251–1263. https://doi.org/10.1177/01461672982412001

Worster, E., Pimperton, H., Ralph-Lewis, A., Monroy, L., Hulme, C., & MacSweeney, M. (2018). Eye movements during visual speech perception in deaf and hearing children. *Language Learning*, *68*(S1), 159–179. https://doi.org/10.1111/lang.12264

Wright, B. D., & Linacre J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370. https://www.rasch.org/rmt/rmt83b.htm

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning*, *61*(4), 1222–1255. https://doi.org/10.1111/j.1467-9922.2011.00667.x

Yan, X., & Chuang, P.-L. (2023). How do raters learn to rate? Many-facet Rasch modeling of rater performance over the course of a rater certification program. *Language Testing*, *40*(1), 153–179. https://doi.org/10.1177/02655322221074913

Yentes, R. D., & Wilhelm, F. (2018) careless: Procedures for computing indices of careless responding. *R packages* [version 1.2.0]. https://github.com/ryentes/careless

Young, R. (2002). Discourse approaches to oral language assessment. *Annual Review of Applied Linguistics*, *22*, 243–262. https://doi.org/10.1017/S0267190502000132

Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hiinkel (Ed.), *Handbook of research in second language teaching and learning volume II* (pp. 426–443). Routledge.

Young, R., & He, A. W. (Eds.). (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Benjamins. https://doi.org/10.1075/sibil.14

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

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, *27*(1), 119–140. https://doi.org/10.1177/0265532209347363

Zhang, Y., & Elder, C. (2011). Judgements of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28*(1), 31–50. https://doi.org/10.1177/0265532209360671

Zhu, X., Liao, X., & Cheong, C. M. (2019). Strategy use in oral communication with competent synthesis and complex interaction. *Journal of Psycholinguistic Research*, *48*, 1163–1183. https://doi.org/10.1007/s10936-019-09651-0

APPENDIX A: INFORMATION, CONSENT, AND NON-DISCLOSURE AGREEMENT

Research Participant Information and Consent Form

Study Title:	Research on Online Language Tests of Speaking
Researcher and Title:	John Dylan Burton, PhD Student
Department and Institution:	Second Language Studies, Michigan State University
Contact Information:	burtonjd@msu.edu, 517-604-1486

BRIEF SUMMARY

I am an PhD student at Michigan State University, currently studying in the Second Language Studies program, and I would like to invite you to take part in a research study about some aspects of speaking tests in an online environment.

Please take time to read the following information carefully before you decide whether or not you wish to take part. Researchers are required to provide a consent form to inform you about the research study, to convey that participation is voluntary, to explain risks and benefits of participation including why you might or might not want to participate, and to empower you to make an informed decision. You should feel free to discuss and ask the researchers any questions you may have.

You are being asked to participate in a research study of speaking tests in an online environment. Because most speaking tests take place face-to-face, we would like to understand more about how people perceive online tests. Your participation in this study will take about two hours total. You will be asked to learn how to use a rating scale and apply it to 10 speaking tests online. Each test will last about 5 minutes. After this, you will be asked a few final follow-up questions.

There is no risk in taking part in this study. Your personal details will not be made public, and your anonymity will be maintained.

You will receive compensation outlined below for your participation. Your input will greatly help us understand how raters perceive online tests of speaking and how they assign scores. The information you provide may be used in the future to create better and more interesting tests.

PURPOSE OF RESEARCH

I have approached you because you have been identified as an experienced rater and you have expressed interest in this study. You and other experienced raters will be able to provide important insight for the development of speaking tests, both in the format online and in the scales used to rate. Your experience is valued greatly and your knowledge can contribute to our understanding of this aspect of test development. I would be very grateful if you would agree to take part in this study.

WHAT YOU WILL BE ASKED TO DO

You will be asked to rate 5-minute speaking tests that have taken place via ZOOM. You will be able to rate these from the comfort of your home or office. Although it is preferrable for you to rate the session in one sitting, you may choose to take breaks as you wish. After rating, there will be a series of questions about yourself and your experience rating. I estimate that the session will take 2 hours. You will receive compensation for your participation in this study.

POTENTIAL RISKS
There are no risks in taking part in this study. Your privacy will be kept by ensuring that no personal information is connected to your rating data or interview data, and all records of your participation will be kept separate from any research data.

PRIVACY AND CONFIDENTIALITY
The data for this project are being collected anonymously. Neither the researchers nor anyone else will be able to link data to you, such as your name, IP address, e-mail address, etc.

The data will be kept in an online repository. Your personal information will not be available in this repository. Only your ratings and transcribed interview data will be available in this repository and may be used by myself and/or other researchers for analysis. Researchers at MSU, the Institutional Review Board (IRB), and academics with access to the data repository will have access only to your ratings and spoken data.

Your personal information will be kept safe and secure in an alternate location from any files available for analysis. Your coworkers and students will not be able to access any of your personally identifiable data.

Although we will make every effort to keep your data confidential there are certain times, such as a court order, where we may have to disclose your data.

The results of this study may be published or presented at professional meetings, but the identities of all research participants will remain anonymous.

Your rights to participate, say no, or withdraw
Participation is voluntary. Refusal to participate will involve no penalty or loss of benefits to which you are otherwise entitled. You may discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled.

You have the right to say no. You may change your mind at any time and withdraw. You may choose not to answer specific questions or to stop participating at any time. Choosing not to participate or withdrawing from this study will not impact you.

If you change your mind, you are free to withdraw at any time during your participation in this study. If you want to withdraw, please let me know, and I will attempt to extract any ideas or information (data) you contributed to the study and destroy them. However, this may be impossible as the questionnaires and recorded tests will be anonymous and randomized, so please tell me as early as possible.

COSTS AND COMPENSATION FOR BEING IN THE STUDY
You will receive $50 in compensation for your participation in this study ($25 per hour) in the form of an Amazon gift card. You will be compensated upon completion of your study.

future research
Information that identifies you will be removed from all data files. The data could be used for future research studies or distributed to another investigator for future research studies without additional informed consent from you.

Contact Information
If you have concerns or questions about this study, such as scientific issues, how to do any part of it, or to report an injury, please contact the researcher:

Name:	Dylan Burton
Mailing Address:	619 Red Cedar Road Michigan State University East Lansing, MI 48824
E-mail address:	burtonjd@msu.edu
Phone number:	517-604-1486

If you have questions or concerns about your role and rights as a research participant, would like to obtain information or offer input, or would like to register a complaint about this study, you may contact, anonymously if you wish, the Michigan State University's Human Research Protection Program at 517-355-2180, Fax 517-432-4503, or e-mail irb@msu.edu or regular mail at 4000 Collins Rd, Suite 136, Lansing, MI 48910.

[RATING STUDY ONLY] Documentation of Informed consent

I agree to allow quotes from the written comments, but not my personal information, to be disclosed in reports and presentations.

☐ *Yes* ☐ *No* Initials_____

Your signature below means that you voluntarily agree to participate in this research study.

Signature _____ Date_____

If you agree to participate in this study, please tick the above boxes, type your initials, and insert an electronic signature in the space above, and write the date. Please save a copy of this consent form on your computer for your own records.

[STIMULATED RECALL ONLY] Documentation of Informed Consent

Participation in this study requires that the interviews be audio recorded. Audio will be transcribed. The audio will not be available to external researchers, but transcriptions will.

I agree to allow audiotaping of the interview.

☐ *Yes* ☐ *No* Initials_____

I agree to allow quotes from the audio transcript, but not my personal information, to be disclosed in reports and presentations.

☐ *Yes* ☐ *No* Initials_____

Your signature below means that you voluntarily agree to participate in this research study.

Signature _____ Date_____

Non-disclosure agreement (NDA)

This Nondisclosure Agreement or ("Agreement") has been entered into on the date of December 1, 2021 and is by and between:
Party Disclosing Information: John Dylan Burton with a mailing address of 619 Red Cedar Road, Michigan State University, East Lansing, MI 48824 (“Disclosing Party”).
Party Receiving Information: {Examiner’s name automatically populated} with a contact address of {populate e-mail address} (“Receiving Party”).
For the purpose of preventing the unauthorized disclosure of Confidential Information as defined below. The parties agree to enter into a confidential relationship concerning the disclosure of certain proprietary and confidential information ("Confidential Information").
1. **Definition of Confidential Information**. For purposes of this Agreement, "Confidential Information" shall include all information or material that has or could have commercial value or other utility in the business in which Disclosing Party is engaged. All audiovisual content (both videos, test questions, and responses) constitute Confidential Information in the context of this research.
2. **Exclusions from Confidential Information**. Receiving Party's obligations under this Agreement do not extend to information that is: (a) publicly known at the time of disclosure or subsequently becomes publicly known through no fault of the Receiving Party; (b) discovered or created by the Receiving Party before disclosure by Disclosing Party; (c) learned by the Receiving Party through legitimate means other than from the Disclosing Party or Disclosing Party's representatives; or (d) is disclosed by Receiving Party with Disclosing Party's prior written approval.
3. **Obligations of Receiving Party**. Receiving Party shall hold and maintain the Confidential Information in strictest confidence for the sole and exclusive benefit of the Disclosing Party. Receiving Party shall not allow access to Confidential Information to any other individuals. Receiving Party shall not, without the prior written approval of Disclosing Party, use for Receiving Party's benefit, publish, copy, or otherwise disclose to others, or permit the use by others for their benefit or to the detriment of Disclosing Party, any Confidential Information. Receiving Party shall return to Disclosing Party any and all records, notes, and other written, printed, or tangible materials in its possession pertaining to Confidential Information immediately if Disclosing Party requests it in writing.
4. **Time Periods**. The nondisclosure provisions of this Agreement shall survive the termination of this Agreement and Receiving Party's duty to hold Confidential Information in confidence shall remain in effect until the Confidential Information no longer qualifies as a trade secret or until Disclosing Party sends Receiving Party written notice releasing Receiving Party from this Agreement, whichever occurs first.
5. **Relationships**. Nothing contained in this Agreement shall be deemed to constitute either party a partner, joint venture or employee of the other party for any purpose.
6. **Severability**. If a court finds any provision of this Agreement invalid or unenforceable, the remainder of this Agreement shall be interpreted so as best to affect the intent of the parties.
7. **Integration**. This Agreement expresses the complete understanding of the parties with respect to the subject matter and supersedes all prior proposals, agreements, representations, and understandings. This Agreement may not be amended except in writing signed by both parties.
8. **Waiver**. The failure to exercise any right provided in this Agreement shall not be a waiver of prior or subsequent rights.
9. **Notice of Immunity**. Receiving Party is provided notice that an individual shall not be held criminally or civilly liable under any federal or state trade secret law for the disclosure of a trade secret that is made (i) in confidence to a federal, state, or local government official, either directly or indirectly, or to an attorney; and (ii) solely for the purpose of reporting or investigating a suspected violation of law; or is made in a complaint or other document filed in a lawsuit or other proceeding, if such filing is made under seal. An individual who files a lawsuit for retaliation by an employer for reporting a suspected violation of law may disclose the trade secret to the attorney of the individual and use the trade secret information in the court proceeding, if the individual (i) files any document containing the trade secret under seal; and

(ii) does not disclose the trade secret, except pursuant to court order.
This Agreement and each party's obligations shall be binding on the representatives, assigns and successors of such party. Each party has signed this Agreement through its authorized representative.

DISCLOSING PARTY

Signature:
Typed or Printed Name:

Date:

RECEIVING PARTY

I hereby agree not to disclose any confidential information as outlined in this agreement, and agree that my initials will be taken as a digital signature.
☐ *Yes* ☐ *No* Initials_____

Full Name:

Date: _____

APPENDIX B: RATING STUDY SIGN-UP, INSTRUCTIONS, AND PRACTICE

Sign-up questionnaire

Thank you for your interest in taking part in this study on foreign language testing.

In order to take part in this study, you must:

- Be an undergraduate at MSU
- Speak English as your first language
- Use a laptop or desktop computer to do the study (no mobile devices)

If you are selected for the study, you will watch several videos and answer questions about them. If you choose to do the study in our lab, participation will take about **two hours on one day**. If you choose to do the study online, participation in the project will take place on **two days**, and each session will last about **one hour (two hours total)**. You may only take part in this study **once**.

You will need:

- **two hours**
- **a quiet place**, **free from distractions**
- **headphones to listen to the audio if you can.**

You can choose when and where you do the study, but you must make sure that you have enough time to complete each session the study in one sitting before you begin. You can take breaks while doing the study,

As compensation, after you complete both sessions, you will receive a $30 e-gift certificate from Amazon.

Click here to read more information about the study. This document also contains the non-disclosure agreement you must agree to prior to beginning the study.

If you have any questions about the study, you may contact me first at burtonjd@msu.edu.

First, I would like to know a little about you

First name: _____

Last name: _____

E-mail address: _____
(E-mail Address: This must be an @msu.edu address, used to send you the survey and gift card at the end.)

Year of birth: _____

Gender (choose): ○Male ○Female ○Other ○I would prefer not to say

Country of Origin/Nationality: _____

Do you speak more than one language? ○Yes ○No

What do you consider your first language? ○English ○Spanish ○Chinese ○Other
(This refers to the language that you grew up speaking and that you use every day. If you are bilingual, it is the language you feel is most dominate.)

What year are you in at MSU? ○Freshman ○Sophomore ○Junior ○Senior
○Other: _____ (please specify)

What is your major? _____

Do you have access to a quiet, distraction-free space, where you can complete this study?
○Yes ○No

After completing the study, a small number of participants may be asked to take part in a face-to-face interview to discuss their scores. Interviewees will be compensated for their time. Would you be willing to meet with the researcher in a lab in Wells Hall to discuss your scores?
○Yes ○No

Thank you, ${q://QID255/ChoiceTextEntryValue}! I will be in touch with you shortly if you are a good fit for the study. If you have any questions, you can write to me at burtonjd@msu.edu.

Figure B.1
Instructions and Practice

Instructions and Practice

In this study, you will **watch** and **listen to** and score students taking a test of English as a foreign language. You will rate them on a series of language- and person-related qualities.

The language qualities you will rate are:

- Fluency: Fluent/Disfluent (rate of speech, breakdowns, and repair)
- Vocabulary: Weak/Strong (range, accuracy, and complexity of words)
- Grammar: Weak/Strong (range, accuracy, and complexity of grammar)
- Pronunciation: Incomprehensible/Comprehensible (how difficult/easy it is to understand the person)

The character qualities you will rate are:

- Engagement (with the test situation)
- Anxiety
- Confidence
- Warmth
- Attentiveness
- Expressiveness
- Happiness
- Competence
- Interactivity (with the examiner)
- Attitude

Think carefully about what these words mean to you. This will be important when you are rating the videos.

You will use a scale with opposite endpoints. You can drag the scale along seven points to choose how closely you feel the person matches with the description. For example, for vocabulary, if you feel the speaker is very strong, you can move the slider to the extreme end:

If the opposite is true, you can move the slider in the other direction:

If you feel the person lies somewhere between weak and strong, you can select a midpoint:

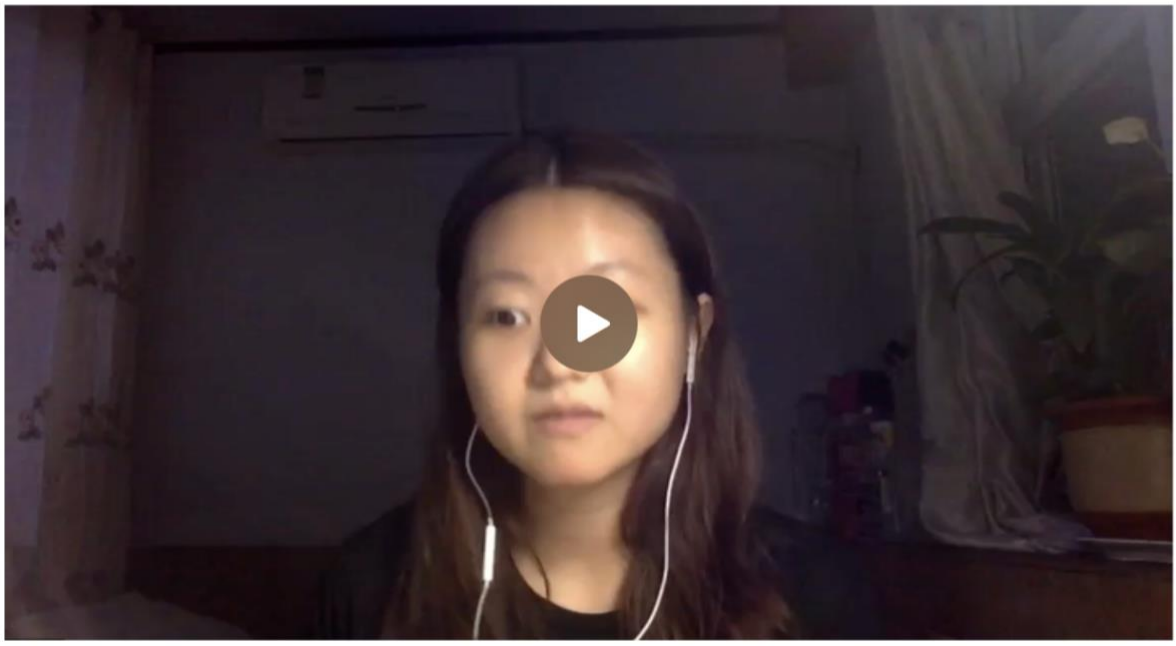
In order to submit a response, you must move the slider first, even if you choose to select the midpoint.

Next

Figure B.2
Practice Set 1

Let's practice before we begin. You are going to watch two practice videos and score them. You can watch the videos multiple times if you wish, but in the main research study you can only watch each video once.

What do you think about this speaker's language and behavior? What scores would you give her on fluency, grammar, vocabulary, and pronunciation?



*Proprietary speech sample, participant signed release

Figure B.3
Practice Set Responses

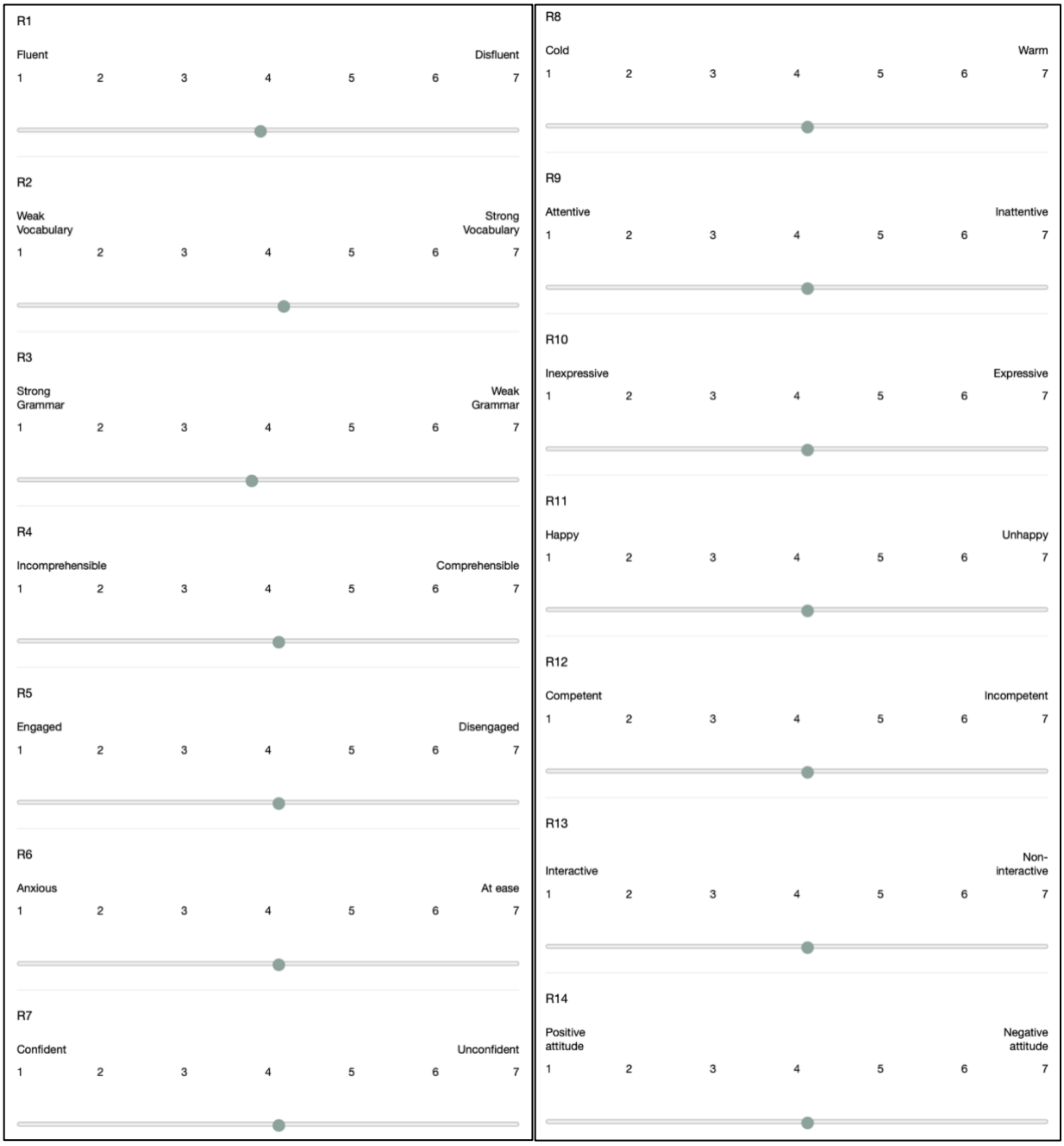


Figure B.4
Practice Set 2

Although the speaker struggles to understand the question at first, overall her language is fairly strong once she begins speaking. She manages to communicate fairly effectively.

Compare her performance with the following example. What score would you give this test-taker?

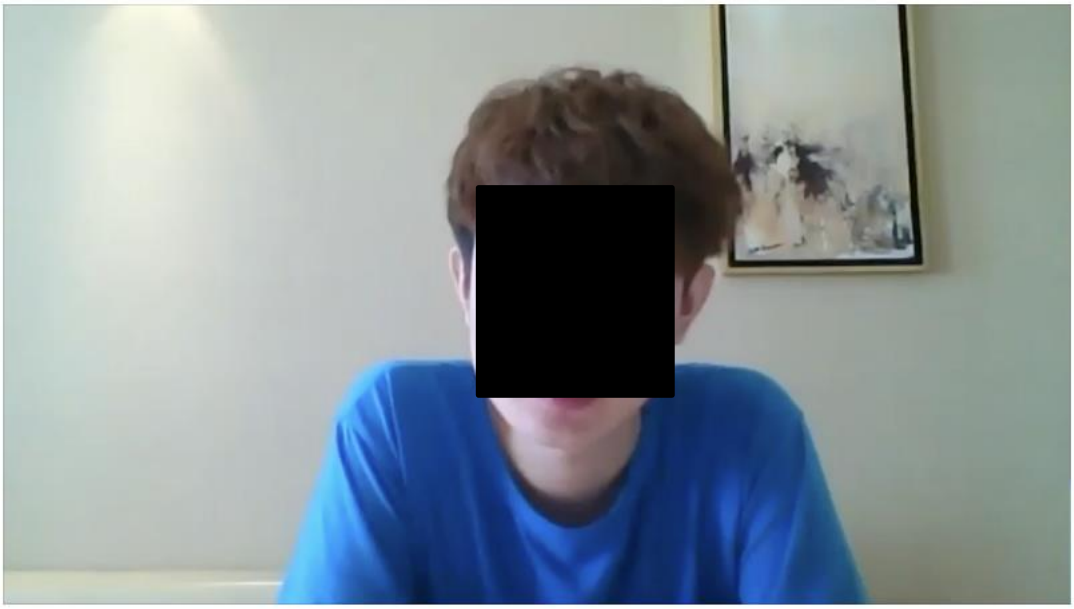
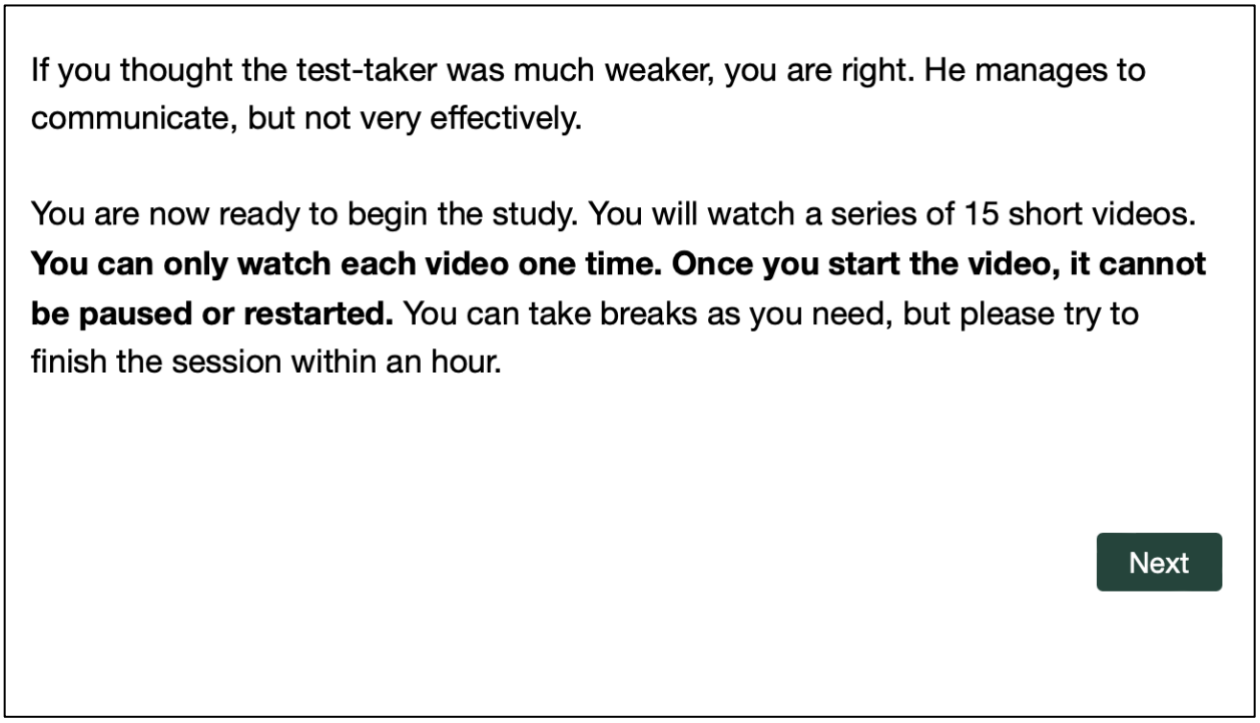


*image removed to protect test taker's identity, but was visible to participants who signed NDAs

[**ABOVE SCALES ARE PRESENTED AGAIN FOR THIS TEST TAKER]**

Figure B.5
Feedback Example



APPENDIX C: FOLLOW-UP SURVEY

Thank you, ${e://Field/First%20Name}! You have now completed rating all samples. I would like to ask you just a few more questions before we finish.

Please indicate your agreement with the following statements:

	Fully agree	Agree	Not sure	Disagree	Fully disagree
The instructions were clear and effective					
The practice sessions were useful					
Rating in the online system was comfortable					
The audio and video were clear					
The rating scales were simple to use					
I feel confident about the ratings I awarded					

If you experienced any technical problems while rating, please let me know here:
_____

If you experienced any doubts about your scores, what made you feel this way? How did you resolve these doubts?
_____

Which rating criterion was the hardest to apply? Why? _____

Please rank the following features for how important they were **when making a decision about language scores** (fluency, grammar, vocabulary, pronunciation). Click and drag the elements; 1 is the most important, 10 is the least.

- The speakers' environment
- The quality of the video and audio
- Mistakes speakers made
- The things speakers talked about (content)
- The speakers facial reactions during the test
- The speakers' appearance (clothing, makeup)
- The examiner (questions and speaking style)
- The speakers' eye gaze and attention
- The speakers' accent
- The words and expressions speakers used

Please rank the following features for how important they were **when making a decision about affect scores** (confidence, anxiety, engagement, etc). Click and drag the elements; 1 is the most important, 10 is the least.

- The speakers' environment
- The quality of the video and audio
- Mistakes speakers made
- The things speakers talked about (content)
- The speakers facial reactions during the test
- The speakers' appearance (clothing, makeup)
- The examiner (questions and speaking style)
- The speakers' eye gaze and attention
- The speakers' accent
- The words and expressions speakers used

Thank you, ${q://QID255/ChoiceTextEntryValue}! I will be in touch soon with your Amazon e-gift certificate.

APPENDIX D: STIMULATED RECALL MATERIALS

Set-up Instructions

Prior to coming to the interview, each participant will be invited to the interview and will select a day and time on the main researcher's schedule using Calendly. Three days prior to the session, each participant will receive an e-mail providing instructions about how to take part in the study. Participants will be told that they will complete the survey online within 24 hours of the stimulated recall. On the day prior to the interview, participants will be e-mailed a link to the survey and instructed to complete it before the interview session at a time convenient to them

The main researcher will carry out the stimulated recall alone without research assistants. The recall will be done with one participant at a time. Each participant will view 10 videos, provide their recalls, and then conduct a wrap-up interview. The entire session will take no longer than 90 minutes, and participants will be compensated for their time.

Recording devices

The main researcher will record sessions using both Quicktime and a digital recording device. Quicktime will ensure video information about stopping points is available for analysis.

Initiation

- Welcome participant to the lab
- Invite them to sit at a computer terminal to the left of mine
- Provide a bottle of water
- Engage in small talk before getting started
- Have the participant sign a consent form

Instructions for research participants

What we're going to do now is re-watch 10 of the videos from the testing survey. I am interested in what you were thinking at the time you were watching and making decisions. So what I'd like you to do is tell me what you were thinking, what was in your mind ***at that time****. Any thoughts you were thinking that you can remember are important.*

We are going to watch the videos on this computer screen. You can pause the video any time you like by pressing the space bar. So if you remember something that you thought when you saw the video, you can push pause and tell me. I might want to ask about what you were thinking in a particular point, in which case I will pause the audio and ask you to talk about your thought processes at that moment.

Now we are going to practice. Here is one of the sample videos that we will watch. To play the video, you will press the space bar.

Demonstrate pressing the space bar. Allow the video to progress about 10 seconds.

Now imagine here you remember thinking something. You would hit the space bar again to pause the video.

Press the space bar to pause the video.

At this point you can tell me what you remember thinking when you watched the video the first time. Take as much time as you need. When you are finished, you can hit the space bar to continue.

Press the space bar to continue. Stop the video.

I may also choose to ask you a question at a particular moment. In that case, I will pause the video myself and ask you a question. Do you have any questions about how this method will work? If you need to take a break at any time just let me know.

Ask if the participant is ready to begin. If the participant is ready, begin recording the session, and allow the participant to start the first video when they are ready.

Probe questions

What were you thinking here/at this point/right then?

What were your first impressions at this point?

What about here/this point? What were you thinking?

Do you remember thinking anything at this point in the video?

Can you tell me what you were thinking at that point?

Can you tell me what you thought when she said that?

How did you arrive at this decision?

Could you elaborate a bit more on that/this/this particular point?

Can you talk about this point a bit more?

Do you have any other comments on this file?

Do you remember anything else you thought about this video?

What makes you think that?

Final interview

After all samples have been finished, conclude with some final questions.

I was wondering if I could ask you something now that the videos are done? I noticed that sometimes you mentioned how the speaker behaved when you were making your decisions about language.

- *How did the person's overall behavior influence your scores? Did anything specific give you more information?*
- *What about eye gaze/facial behaviors/posture?*
- *How do you balance how they behave with the things that they say when you make decisions?*
- *Do you think any specific aspect of language is impacted by nonverbal behavior more than others?*

APPENDIX E: COMMUNICATIONS TO PARTICIPANTS

Sign-up letter

Dear MSU students,

I am excited to invite you to participate in my research study on online foreign language speaking tests. This project is part of my dissertation. The goal of the study is to understand how people perceive second language English speech of individuals taking an online speaking test. By participating, you will help us better understand how the online environment impacts language processing.

Who can participate?
Up to 60 undergraduates at Michigan State University who are first language English speakers

What will you do?
You will watch and listen to a set of speech samples online and rate them using a set of rating scales.

How long will it take?
Approximately 2 hours spread out on two different days, online

What will you get for participating?
An Amazon gift card for $30 USD

Why should you participate?
To contribute to our understanding of second language speech perception and to help improve the online formats of speaking tests

If you are interested in participating, please go to https://msu.co1.qualtrics.com/jfe/form/SV_8faVZhqlS8FzGB0 to fill out a screener in order to participate. Participation is first come first serve and will be subject to meeting the requirements stipulated above. Please do not hesitate to contact me with any questions regarding your participation.

Best,
Dylan Burton

Invitation letter for day 1

Dear ${e://Field/First%20Name},

Thank you for your recent expression of interest in participating in a research study on foreign language testing. I have reviewed your survey response and believe you would be a good fit for the study.

Before starting:

- Please remember to review the study's information sheet and non-disclosure agreement. You will be asked to digitally agree to these before beginning the study.
- You will also need to choose a day and a time at your own convenience when you can complete the study in one sitting.
- Each of the two sessions will take about an hour, and as you will be watching and listening to videos, you will need a quiet space free from distractions. I can arrange time in a computer lab in Wells Hall if you do not have access to a quiet space. Just reply to this e-mail and let me know.
- The study can only be completed on a laptop or desktop computer. No mobile devices are allowed.
- Please try to make time to complete the study by ${date://OtherDate/FL/+1%20week}. If you would like to drop out of the study, please let me know by replying to this e-mail address. There are few places remaining in the study, and once these have been filled, access to the study will be closed.
- Once you complete both days of the study, you will receive a $30 Amazon e-gift card.

When you are ready to begin, you may click the link below. If for whatever reason your window closes while doing the study, you can return to it by using the same link.

Start the survey!

If the above link asks you for "ExternalReferenceID", please enter the code: ${e://Field/id}

If you have any questions please let me know.

Best wishes,
Dylan Burton
PhD Candidate, Second Language Studies

${l://OptOutLink}
${l://SurveyLink?d=Take%20the%20survey}

Invitation letter for day 2

Dear ${e://Field/First%20Name},

Thank you for completing Day 1 of the research study on foreign language tests. You may now complete the second set of speech samples by clicking the below link. You can do this at the day/time of your choosing, just like Day 1 of the study. Remember to set aside one hour to complete the study, and make sure you have a quiet space free from distractions before you start.

You can access Day 2 of the study below. If your window closes during the study, you can return to it using the same link.

https://msu.co1.qualtrics.com/jfe/form/SV_eXmSBnjjkXnlgwK?id=${e://Field/id}

If the link above asks you for a code, please enter ${e://Field/id}

Please let me know if you have any problems or questions.

Best wishes,
Dylan Burton

Invitation letter for study and stimulated recall

Dear [Name],

Thank you for expressing interest in taking part in the research study on foreign language testing. I have looked over your response, and I believe you would be a good fit for the study.

In the survey, you indicated that you would be willing to be interviewed face-to-face following completion of the study in our lab in Wells Hall. If this is still the case, I would like to offer you the chance to take part in the study. I will in turn give you an Amazon e-gift card worth $50 for participating: $30 for the study itself, and $20 for the interview.

In order to take part, first choose a 90-minute spot for the face-to-face interview using the following link: https://calendly.com/burtonjd/foreign-language-testing-research-study. 24-hours before the interview, I will send you a link to complete the study in your home or other quiet place. The study will take about two hours to complete. You can do this on the same day as the interview, but you must have enough time to finish before the interview starts. If you wish to do the study in our lab prior to the interview, this can also be arranged. Just let me know.

If you are unable to participate, or if you are no longer interested, please just let me know so that I can invite someone else in your place.

Thank you so much. I look forward to hearing from you.

Best,
Dylan Burton
PhD Candidate, Second Language Studies

Instructions e-mail for stimulated recall

Dear [Name],

Thank you for agreeing to participate in the research study on foreign language testing! This e-mail contains instructions on how the study will proceed.

Instructions:

- On [Date] at [Time], I will send you a link to the survey. You will need to choose a time within 24 hours of our interview start time to complete this study online. It must be fully complete by the time our interview starts, so make sure you plan accordingly.
- The survey will take about two hours and it includes a break. As you will be watching and listening to videos, you will need a quiet space free from distractions during that time.
- You can use our lab in Wells Hall if you do not have access to a quiet space. If you would like, this can be just before our interview. Just reply to this e-mail and let me know and I can schedule this for you.
- The study can only be completed on a laptop or desktop computer. No mobile devices are allowed.
- We will meet in Wells Hall in room B417 (B-wing) on [Date] at [Time] for a 90-minute interview. You do not need to prepare anything for the interview, but please arrive on time.
- Once you complete the study and the interview, you will receive a $50 Amazon e-gift card.

If you cannot complete both the survey and the interview, or if you have any other last minute changes or requests, please let me know by replying to this e-mail address or contacting me at 517-604-1486.

Thank you!
Dylan Burton
PhD Candidate, Second Language Studies

Invitation letter for one-day rating study

Dear ${e://Field/First%20Name},

Thank you for agreeing to participate in the research study on foreign language testing. This e-mail contains the link to the survey, and following this we will meet in Wells Hall for a face-to-face interview.

Before starting:

- Please remember to review the study's information sheet and non-disclosure agreement. You will be asked to digitally agree to these before beginning the study.
- You will need to choose a time within 24 hours of our interview start time to complete this study. It must be fully complete by the time our interview starts, so make sure you plan at least two hours to complete the survey beforehand.
- The survey will take about two hours and it includes a break. As you will be watching and listening to videos, you will need a quiet space free from distractions. You can use our lab in Wells Hall if you do not have access to a quiet space. If you would like, this can be just before our interview. Just reply to this e-mail and let me know and I can schedule this for you.
- The study can only be completed on a laptop or desktop computer. No mobile devices are allowed.
- If you cannot complete both the survey and the interview, or if you have any other last minute changes or requests, please let me know by replying to this e-mail address or contacting me at 517-604-1486.
- Once you complete the study and the interview, you will receive a $50 Amazon e-gift card.

When you are ready to begin, you may click the link below. If for whatever reason your window closes while doing the study, you can return to it by using the same link.

${l://SurveyLink?d=Take%20the%20survey}

If the above link asks you for "ExternalReferenceID", please enter the code: ${e://Field/id}

If you have any questions please let me know.

Best wishes,
Dylan Burton
PhD Candidate, Second Language Studies

${l://OptOutLink}

APPENDIX F: ELAN TIER DESCRIPTIONS (ADAPTED FROM BURTON, 2021)

Table F.1

ELAN Tier Descriptions

Tier	Label	Description
1-2	*Discourse*	The examiner's and the participants' speech were separately transcribed into two tiers. In most cases these represented full TCU. These units were segmented through a frame-by-frame analysis of the audio and an inspection of the waveform output. These were transcribed orthographically with breathing, filled pauses, and laughing indicated.
3-4	*Words*	The examiner and the participants' speech were likewise separately transcribed word by word using orthographic transcription. Word boundaries were segmented by use of the waveform and a reduced speed audio recording. Breathing was transcribed using <.hhh> for inhaling and <hhh.> for exhaling. Filled pauses were marked with "uh" or "um" as spoken by the test taker. Laughing was annotated as <hhuh>.
5	*Gaze*	Gaze was annotated as *averted* from the moment of the first shift in eye direction until gaze was reconnected with the examiner. When blinking occurred at gaze shift boundaries, the blink was included with averted gaze.
6	*Blinks*	Blinks were segmented separately from gaze. Blink segments were annotated from the first moment the participant's eyelid began to fall and ended when the eyeball again became visible.
7	*Mouth*	Three mouth behaviors were annotated. *Pursed Lips* were annotated when the participants' mouth was tightly closed, generally with the cheek muscles tight on each side of the lips. *Smiling* was annotated without distinguishing Duchenne and non-Duchenne types. *Open (non-speaking)* was a category that appeared between speech segments where the participant held her mouth open without speech. *Laughing* was annotated only when this behavior was seen and heard. *Tongue touching lips* was annotated when the mouth was closed with the tongue visible.
8	*Eyebrow*	Two eyebrow movements were found in the dataset. *Furrowed* brows were contracted, often with visible skin folds between the eyebrows. *Raised* indicated eyebrows lifted vertically away from the eyes.
9	*Head Turn*	Head movements were classified as *head turn left, head turn right, head tilt left, head tilt right, head raise,* and *head lower*. These behaviors described movement when the head moved either to the left or right (*turn*), in a diagonal direction (*tilt*), or the specified vertical direction. Turns, raises, and lowers generally accompanied averted gaze.
10	*Head Gesture*	Head gestures included *nods, shakes,* and *pokes*. *Nods* included nonverbal backchannels and were annotated throughout their duration. *Shakes* were annotated as moments when an individual disagreed or negated a statement with the head turning side to side quickly. *Pokes* were annotated as a quick head movement forward that may convey nonunderstanding (Seo & Koshik, 2010).
11	*Posture*	Posture referred broadly to the relationship between the participants' body and the camera. *Tilt forward* occurred when the participant leaned from the neutral position towards the camera. *Tilt back* referred to movement away from the camera from the neutral position. *Rocking* was annotated when the test taker was seen moving backward and forward in relation to the camera. *Shift* was annotated when an individual quickly moved right or left in their chair to readjust.
12	*Gesture*	Gestures were segmented broadly as general movements with the hand. *Representational gestures* are gestures occurring with speech with a non-emblematic visual or metaphorical referential meaning (Kendon, 2004) and were annotated by describing them as closely as possible. *Deictics* were annotated when an individual pointed. *Beats* were annotated when gestures were used to emphasize speech at prosodic boundaries. Self-*adaptors* are movements, generally of the hands, which may not co-occur with speech and generally are not representational in meaning (Ekman and Friesen, 1969). These were annotated as a description of the action taking place (e.g., scratches head).
13	*Other*	This category was left for occasional movements that were rarely observed in the dataset. *Swallowing* and *shoulder shrugging* were annotated when visible.

APPENDIX G: STUDY VARIABLES

Table G.1
Summary of Variables

Variable	Definition	Type	Format
Background variables (not modeled)			
Age	The rater's age at time of participation	Continuous/Ratio	Integer
Gender	The gender that a rater identifies with	Categorical	Four categories: Male/Female/Other/Prefer not to say
Nationality	The raters' country of origin	Categorical	Open-ended/Rater may specify
L1	The raters' first language or language considered most dominant	Categorical	Open-ended/Rater may specify
L2	Yes/no question of whether the participant spoke an L2	Binary choice	Yes/No
Major	Main focus of study in college education	Categorical	Open-ended/Rater may specify
Predictor variables			
Valence	Overall positivity/negativity of emotional response	Ratio	Mean score from -100 – 100
Attention	Directedness of eye gaze and head turns to webcam camera	Ratio	Mean score of 0 – 100
Engagement	A measure of overall facial muscle activation (expressiveness)	Ratio	Mean score of 0 – 100
IELTS Test Scores	Test scores originally reported on 1–9 band scale rescaled to 1–7 scale for comparability with other measures in this study	Ranked categorical	Score of 1–7
Affect scales (10)	Semantic differential scales of engagement, anxiety, confidence, warmth, attention, expressiveness, happiness, competence, interactiveness, attitude	Ranked categorical	Score of 1–7
Outcome variables			
Language scales (4)	Semantic differential scales of fluency, grammar, vocabulary, and comprehensibility	Ranked categorical	Score of 1–7

APPENDIX H: CATEGORY STATISTICS

Figure H.1
Rasch Category Statistics

Model = ?,?,1,PROFICIENCY ; Criteria: Fluency
Rating (or partial credit) scale = PROFICIENCY,R7,G,O

DATA Category Counts					QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
			Cum.		Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	82	82	3%	3%	-.80	-.80	.9			(-3.14)		low	low	100%
2	265	265	11%	14%	-.65	-.56	.8	-1.86	.12	-1.53	-2.38	-1.86	-2.12	43%
3	400	400	16%	30%	-.23	-.23	.9	-.81	.07	-.62	-1.00	-.81	-.89	35%
4	247	247	10%	40%	.10	.15	.8	.44	.06	-.05	-.32		-.20	16%
5	637	637	26%	66%	.50	.55	.8	-.60	.05	.55	.23	-.08	.08	38%
6	599	599	24%	90%	.99	.96	.8	.81	.05	1.57	.97	.81	.87	46%
7	260	260	10%	100%	1.59	1.42	.9	2.02	.07	(3.29)	2.50	2.02	2.24	100%
										(Mean)		(Modal)	(Median)	

Model = ?,?,2,PROFICIENCY ; Criteria: Vocabulary
Rating (or partial credit) scale = PROFICIENCY,R7,G,O

DATA Category Counts					QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
			Cum.		Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	82	82	3%	3%	-.89	-.87	1.1			(-3.37)		low	low	100%
2	322	322	13%	16%	-.62	-.59	1.0	-2.11	.12	-1.64	-2.57	-2.11	-2.33	46%
3	541	541	22%	38%	-.22	-.23	1.0	-.94	.06	-.56	-1.01	-.94	-.94	41%
4	246	246	10%	48%	.11	.16	.9	.75	.05	.07	-.22		-.05	16%
5	587	587	24%	71%	.53	.55	.9	-.52	.05	.66	.35	.12	.22	37%
6	477	477	19%	91%	.96	.94	.9	.95	.05	1.58	1.04	.95	.96	42%
7	235	235	9%	100%	1.44	1.37	1.0	1.86	.08	(3.17)	2.43	1.86	2.14	100%
										(Mean)		(Modal)	(Median)	

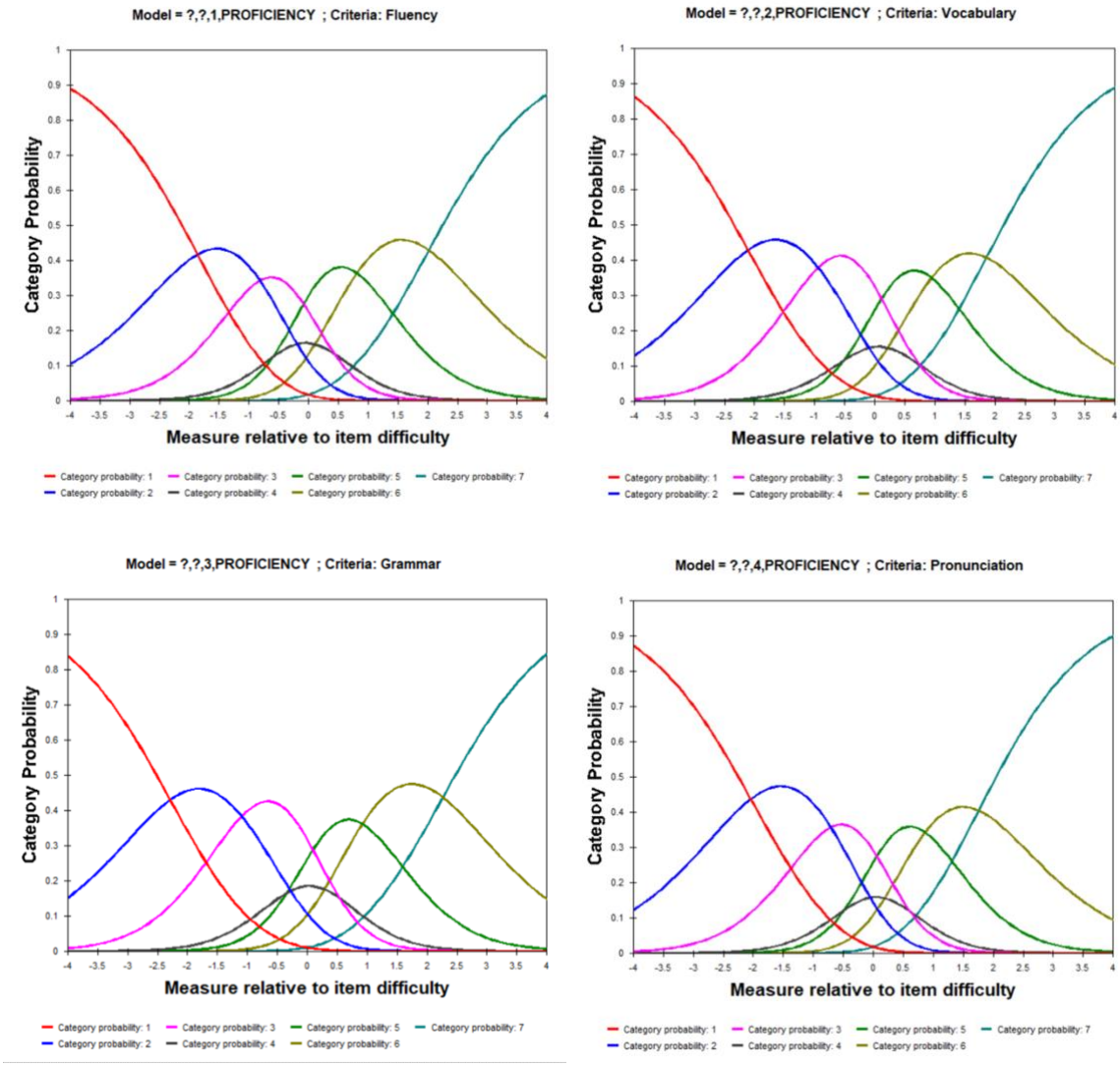
Model = ?,?,3,PROFICIENCY ; Criteria: Grammar
Rating (or partial credit) scale = PROFICIENCY,R7,G,O

DATA Category Counts					QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
			Cum.		Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	83	83	3%	3%	-.91	-1.01	1.4			(-3.54)		low	low	100%
2	342	342	14%	17%	-.61	-.72	1.5	-2.29	.12	-1.80	-2.75	-2.29	-2.51	46%
3	600	600	24%	41%	-.23	-.33	1.3	-1.09	.06	-.66	-1.14	-1.09	-1.09	43%
4	307	307	12%	53%	-.03	.07	1.1	.54	.05	.04	-.28		-.12	19%
5	572	572	23%	76%	.35	.47	1.4	-.35	.05	.69	.35	.10	.22	37%
6	447	447	18%	94%	.86	.88	1.2	.92	.06	1.75	1.13	.92	1.02	48%
7	139	139	6%	100%	1.37	1.32	1.0	2.27	.09	(3.51)	2.70	2.27	2.46	100%
										(Mean)		(Modal)	(Median)	

Model = ?,?,4,PROFICIENCY ; Criteria: Pronunciation
Rating (or partial credit) scale = PROFICIENCY,R7,G,O

DATA Category Counts					QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
			Cum.		Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	49	49	2%	2%	-.37	-.60	1.4			(-3.28)		low	low	100%
2	230	230	9%	11%	-.31	-.37	1.2	-2.04	.15	-1.53	-2.47	-2.04	-2.24	47%
3	370	370	15%	26%	-.02	-.07	1.2	-.71	.07	-.53	-.94	-.71	-.82	37%
4	224	224	9%	35%	.16	.30	.9	.61	.06	.05	-.22		-.08	16%
5	598	598	24%	59%	.60	.69	1.1	-.49	.06	.62	.32	.06	.19	36%
6	617	617	25%	84%	1.08	1.11	.9	.87	.05	1.51	.98	.87	.90	42%
7	402	402	16%	100%	1.68	1.56	.9	1.76	.06	(3.08)	2.33	1.76	2.04	100%
										(Mean)		(Modal)	(Median)	

Figure H.2
Rasch Item Characteristic Curves



APPENDIX I: DESCRIPTIVE DATA FOR RATERS

Figure I.1

Rasch Rater Data Organized by Fit Statistics

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
505	120	4.21	4.25	-.17	.08	1.92	5.8	2.11	6.5	.10	.44	.70	22.6	25.7	10	10
544	120	4.53	4.67	-.41	.08	1.94	5.8	1.95	5.7	.22	.30	.70	20.4	26.3	15	15
541	120	4.51	4.64	-.39	.08	1.64	4.2	1.70	4.4	.55	.57	.70	24.1	26.2	63	63
505	120	4.21	4.25	-.17	.08	1.57	3.9	1.65	4.2	.40	.71	.70	23.5	25.7	11	11
477	120	3.97	3.94	.01	.08	1.47	3.4	1.65	4.2	.56	.58	.70	24.1	24.8	81	81
463	120	3.86	3.79	.09	.08	1.40	2.9	1.64	4.2	.33	.51	.70	20.9	24.2	26	26
579	120	4.82	5.02	-.64	.08	1.64	4.1	1.51	3.3	.58	.69	.69	23.9	26.1	76	76
561	120	4.68	4.84	-.52	.08	1.55	3.7	1.47	3.1	.59	.71	.69	25.1	26.3	16	16
569	120	4.74	4.92	-.58	.08	1.42	2.9	1.49	3.2	.73	.55	.69	25.8	26.2	40	40
606	120	5.05	5.27	-.83	.08	1.27	1.9	1.47	3.0	.54	.58	.67	24.1	25.5	30	30
660	120	5.50	5.71	-1.24	.09	1.44	2.8	1.33	2.1	.78	.57	.63	23.8	23.2	68	68
587	120	4.89	5.09	-.70	.08	1.44	3.0	1.36	2.4	.67	.75	.68	24.5	26.0	83	83
594	120	4.95	5.16	-.75	.08	1.27	1.9	1.40	2.6	.74	.62	.68	24.2	25.8	6	6
532	120	4.43	4.54	-.34	.08	1.26	1.9	1.36	2.4	.98	.66	.70	28.2	26.2	77	77
519	120	4.32	4.40	-.26	.08	1.19	1.4	1.35	2.4	.64	.58	.70	24.6	26.0	8	8
570	120	4.75	4.93	-.58	.08	1.30	2.1	1.33	2.3	.54	.73	.69	23.2	26.2	67	67
510	120	4.25	4.30	-.20	.08	1.32	2.3	1.28	2.0	.85	.76	.70	24.4	25.8	37	37
545	120	4.54	4.68	-.42	.08	1.23	1.7	1.30	2.1	.67	.65	.70	24.9	26.3	57	57
522	120	4.35	4.43	-.27	.08	1.24	1.8	1.29	2.1	.95	.61	.70	27.0	26.0	12	12
459	120	3.83	3.74	.12	.08	1.21	1.6	1.28	2.0	.70	.56	.70	22.8	24.0	73	73
521	120	4.34	4.42	-.27	.08	1.22	1.6	1.17	1.2	1.00	.75	.70	27.9	26.0	14	14
457	120	3.81	3.72	.13	.08	1.15	1.2	1.21	1.5	.86	.65	.70	24.5	23.9	28	28
518	120	4.32	4.39	-.25	.08	1.17	1.3	1.20	1.4	.93	.78	.70	27.0	25.9	61	61
522	120	4.35	4.43	-.27	.08	1.19	1.4	1.18	1.3	.88	.79	.70	25.9	26.0	48	48
553	120	4.61	4.76	-.47	.08	1.18	1.3	1.12	.9	.88	.74	.70	25.4	26.3	74	74
452	120	3.77	3.67	.16	.08	1.11	.8	1.18	1.3	1.04	.59	.70	24.6	23.7	66	66
591	120	4.93	5.13	-.73	.08	1.03	.2	1.17	1.2	.96	.70	.68	27.7	25.9	19	19
537	120	4.47	4.59	-.37	.08	1.10	.7	1.17	1.2	.79	.57	.70	23.6	26.2	62	62
576	120	4.80	4.99	-.62	.08	1.04	.3	1.17	1.2	1.01	.72	.69	28.0	26.1	52	52
559	120	4.66	4.82	-.51	.08	1.10	.8	1.16	1.1	.94	.69	.69	27.7	26.3	46	46
544	120	4.53	4.67	-.41	.08	.95	-.3	1.13	1.0	.91	.40	.70	23.9	26.3	25	25
523	120	4.36	4.44	-.28	.08	1.10	.8	1.07	.5	1.01	.71	.70	26.1	26.0	78	78
557	120	4.64	4.80	-.50	.08	.95	-.3	1.10	.7	.85	.72	.69	26.4	26.3	32	32
541	120	4.51	4.64	-.39	.08	1.03	.2	1.10	.7	1.16	.63	.70	27.5	26.2	47	47
633	120	5.28	5.50	-1.03	.09	1.09	.7	.95	-.2	1.08	.75	.65	26.0	24.5	53	53
528	120	4.40	4.50	-.31	.08	.99	.0	1.09	.6	.97	.68	.70	25.8	26.1	34	34
502	120	4.18	4.21	-.15	.08	.94	-.4	1.06	.5	.97	.74	.70	26.1	25.6	41	41
557	120	4.64	4.80	-.50	.08	1.05	.4	1.04	.3	.94	.82	.69	25.8	26.3	9	9
705	120	5.88	6.04	-1.65	.10	1.05	.3	.97	-.1	1.12	.58	.58	21.8	20.3	50	50
486	120	4.05	4.04	-.05	.08	1.03	.2	1.04	.3	.99	.74	.70	25.2	25.1	39	39
424	120	3.53	3.38	.34	.08	1.03	.2	1.01	.1	.99	.82	.69	22.5	22.2	21	21
575	120	4.79	4.98	-.62	.08	1.02	.1	1.01	.1	.99	.80	.69	27.3	26.1	23	23
543	120	4.53	4.66	-.41	.08	.98	-.1	1.00	.0	1.05	.75	.70	28.3	26.2	80	80
476	120	3.97	3.93	.01	.08	.98	-.1	.96	-.2	1.00	.79	.70	25.0	24.7	59	59
492	120	4.10	4.10	-.09	.08	.94	-.4	.95	-.3	1.26	.67	.70	25.9	25.3	13	13
476	120	3.97	3.93	.01	.08	.95	-.3	.94	-.4	1.32	.71	.70	26.8	24.7	2	2
486	120	4.05	4.04	-.05	.08	.93	-.5	1.01	.1	1.15	.78	.70	27.6	25.1	5	5
668	120	5.57	5.77	-1.31	.09	1.01	.1	.93	-.4	1.23	.70	.62	24.9	22.7	82	82
536	120	4.47	4.58	-.36	.08	.92	-.5	.92	-.5	1.28	.78	.70	28.2	26.2	71	71
592	120	4.93	5.14	-.73	.08	.96	-.3	.92	-.6	1.12	.70	.68	27.0	25.9	64	64
558	120	4.65	4.81	-.50	.08	.91	-.6	.96	-.2	1.11	.79	.69	27.6	26.3	20	20
586	120	4.88	5.09	-.69	.08	.91	-.7	.92	-.5	1.01	.76	.68	25.8	26.0	17	17
443	120	3.69	3.57	.22	.08	.89	-.9	.97	-.2	1.23	.63	.70	24.7	23.2	18	18
540	120	4.50	4.63	-.39	.08	.88	-.9	.90	-.7	1.28	.85	.70	29.0	26.2	75	75
443	120	3.69	3.57	.22	.08	.82	-1.4	.86	-1.1	1.11	.76	.70	23.1	23.2	7	7
533	120	4.44	4.55	-.34	.08	.81	-1.5	.85	-1.1	1.30	.73	.70	28.5	26.2	56	56
502	120	4.18	4.21	-.15	.08	.83	-1.4	.80	-1.6	1.32	.80	.70	28.3	25.6	65	65
467	120	3.89	3.83	.07	.08	.79	-1.8	.88	-.9	1.03	.64	.70	22.6	24.4	31	31
655	120	5.46	5.67	-1.20	.09	.77	-1.7	.91	-.6	1.08	.68	.63	24.2	23.5	60	60
592	120	4.93	5.14	-.73	.08	.76	-1.9	.84	-1.1	1.18	.73	.68	27.9	25.9	27	27
617	120	5.14	5.36	-.91	.09	.77	-1.8	.75	-1.9	1.15	.75	.66	26.4	25.1	54	54
482	120	4.02	3.99	-.03	.08	.76	-2.0	.74	-2.1	1.22	.70	.70	25.8	25.0	22	22

Figure I.1 (cont'd)

511	120	4.26	4.31	-.21	.08	.73	-2.3	.75	-2.0	1.20	.64	.70	25.8	25.8	58 58
569	120	4.74	4.92	-.58	.08	.72	-2.3	.75	-2.0	1.30	.80	.69	28.2	26.2	45 45
557	120	4.64	4.80	-.50	.08	.71	-2.4	.73	-2.2	1.29	.78	.69	28.6	26.3	1 1
396	120	3.30	3.10	.52	.08	.71	-2.5	.81	-1.5	.94	.54	.68	18.5	20.4	70 70
514	120	4.28	4.35	-.22	.08	.70	-2.6	.74	-2.1	.95	.66	.70	24.6	25.9	42 42
464	120	3.87	3.80	.09	.08	.70	-2.7	.74	-2.1	1.10	.69	.70	24.2	24.2	44 44
456	120	3.80	3.71	.14	.08	.69	-2.8	.76	-2.0	.54	.65	.70	19.3	23.9	43 43
507	120	4.22	4.27	-.18	.08	.66	-3.1	.66	-2.9	1.23	.76	.70	26.2	25.7	24 24
529	120	4.41	4.51	-.32	.08	.67	-2.9	.65	-3.0	1.58	.71	.70	30.2	26.1	29 29
512	120	4.27	4.32	-.21	.08	.65	-3.2	.75	-2.0	1.08	.71	.70	25.6	25.8	33 33
445	120	3.71	3.59	.20	.08	.64	-3.3	.78	-1.8	1.11	.67	.70	23.2	23.3	35 35
661	120	5.51	5.71	-1.25	.09	.65	-2.9	.64	-2.9	1.31	.71	.63	24.7	23.1	4 4
530	120	4.42	4.52	-.32	.08	.61	-3.5	.63	-3.2	1.32	.76	.70	28.7	26.1	79 79
546	120	4.55	4.69	-.43	.08	.60	-3.6	.60	-3.5	1.42	.76	.70	28.8	26.3	38 38
558	120	4.65	4.81	-.50	.08	.59	-3.7	.59	-3.6	1.46	.84	.69	30.0	26.3	69 69
540	120	4.50	4.63	-.39	.08	.57	-4.0	.66	-2.9	1.11	.68	.70	26.5	26.2	3 3
561	120	4.68	4.84	-.52	.08	.57	-3.9	.56	-3.9	1.46	.78	.69	29.1	26.3	55 55
479	120	3.99	3.96	-.01	.08	.61	-3.6	.56	-4.0	1.52	.79	.70	28.3	24.8	49 49
649	120	5.41	5.62	-1.15	.09	.64	-3.0	.53	-4.1	1.48	.71	.64	27.2	23.8	72 72
567	120	4.72	4.90	-.56	.08	.53	-4.4	.58	-3.7	1.37	.71	.69	28.3	26.2	36 36
558	120	4.65	4.81	-.50	.08	.56	-4.0	.52	-4.4	1.49	.81	.69	29.4	26.3	51 51
536.2	120.0	4.47	4.56	-.38	.08	1.00	-.2	1.04	.1		.69				Mean (Count: 83)
59.0	.0	.49	.59	.39	.00	.31	2.4	.33	2.4		.10				S.D. (Population)
59.3	.0	.49	.59	.40	.00	.31	2.4	.33	2.4		.10				S.D. (Sample)

Model, Populn: RMSE .08 Adj (True) S.D. .39 Separation 4.75 Strata 6.66 Reliability (not inter-rater) .96
Model, Sample: RMSE .08 Adj (True) S.D. .39 Separation 4.78 Strata 6.70 Reliability (not inter-rater) .96
Model, Fixed (all same) chi-squared: 1770.0 d.f.: 82 significance (probability): .00
Model, Random (normal) chi-squared: 78.3 d.f.: 81 significance (probability): .56
Inter-Rater agreement opportunities: 408360 Exact agreements: 104870 = 25.7% Expected: 103317.0 = 25.3%

Figure I.2

Rasch Rater Data Organized by Severity Measure

Table 7.1.1 Raters Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
396	120	3.30	3.10	.52	.08	.71	-2.5	.81	-1.5	.94	.54	.68	18.5	20.4	70	70
424	120	3.53	3.38	.34	.08	1.03	.2	1.01	.1	.99	.82	.69	22.5	22.2	21	21
443	120	3.69	3.57	.22	.08	.82	-1.4	.86	-1.1	1.11	.76	.70	23.1	23.2	7	7
443	120	3.69	3.57	.22	.08	.89	-.9	.97	-.2	1.23	.63	.70	24.7	23.2	18	18
445	120	3.71	3.59	.20	.08	.64	-3.3	.78	-1.8	1.11	.67	.70	23.2	23.3	35	35
452	120	3.77	3.67	.16	.08	1.11	.8	1.18	1.3	1.04	.59	.70	24.6	23.7	66	66
456	120	3.80	3.71	.14	.08	.69	-2.8	.76	-2.0	.54	.65	.70	19.3	23.9	43	43
457	120	3.81	3.72	.13	.08	1.15	1.2	1.21	1.5	.86	.65	.70	24.5	23.9	28	28
459	120	3.83	3.74	.12	.08	1.21	1.6	1.28	2.0	.70	.56	.70	22.8	24.0	73	73
463	120	3.86	3.79	.09	.08	1.40	2.9	1.64	4.2	.33	.51	.70	20.9	24.2	26	26
464	120	3.87	3.80	.09	.08	.70	-2.7	.74	-2.1	1.10	.69	.70	24.2	24.2	44	44
467	120	3.89	3.83	.07	.08	.79	-1.8	.88	-.9	1.03	.64	.70	22.6	24.4	31	31
476	120	3.97	3.93	.01	.08	.95	-.3	.94	-.4	1.32	.71	.70	26.8	24.7	2	2
476	120	3.97	3.93	.01	.08	.98	-.1	.96	-.2	1.00	.79	.70	25.0	24.7	59	59
477	120	3.97	3.94	.01	.08	1.47	3.4	1.65	4.2	.56	.58	.70	24.1	24.8	81	81
479	120	3.99	3.96	-.01	.08	.61	-3.6	.56	-4.0	1.52	.79	.70	28.3	24.8	49	49
482	120	4.02	3.99	-.03	.08	.76	-2.0	.74	-2.1	1.22	.70	.70	25.8	25.0	22	22
486	120	4.05	4.04	-.05	.08	.93	-.5	1.01	.1	1.15	.78	.70	27.6	25.1	5	5
486	120	4.05	4.04	-.05	.08	1.03	.2	1.04	.3	.99	.74	.70	25.2	25.1	39	39
492	120	4.10	4.10	-.09	.08	.94	-.4	.95	-.3	1.26	.67	.70	25.9	25.3	13	13
502	120	4.18	4.21	-.15	.08	.94	-.4	1.06	.5	.97	.74	.70	26.1	25.6	41	41
502	120	4.18	4.21	-.15	.08	.83	-1.4	.80	-1.6	1.32	.80	.70	28.3	25.6	65	65
505	120	4.21	4.25	-.17	.08	1.92	5.8	2.11	6.5	.10	.44	.70	22.6	25.7	10	10
505	120	4.21	4.25	-.17	.08	1.57	3.9	1.65	4.2	.40	.71	.70	23.5	25.7	11	11
507	120	4.22	4.27	-.18	.08	.66	-3.1	.66	-2.9	1.23	.76	.70	26.2	25.7	24	24
510	120	4.25	4.30	-.20	.08	1.32	2.3	1.28	2.0	.85	.76	.70	24.4	25.8	37	37
511	120	4.26	4.31	-.21	.08	.73	-2.3	.75	-2.0	1.20	.64	.70	25.8	25.8	58	58
512	120	4.27	4.32	-.21	.08	.65	-3.2	.75	-2.0	1.08	.71	.70	25.6	25.8	33	33
514	120	4.28	4.35	-.22	.08	.70	-2.6	.74	-2.1	.95	.66	.70	24.6	25.9	42	42
518	120	4.32	4.39	-.25	.08	1.17	1.3	1.20	1.4	.93	.78	.70	27.0	25.9	61	61
519	120	4.32	4.40	-.26	.08	1.19	1.4	1.35	2.4	.64	.58	.70	24.6	26.0	8	8

Figure I.2 (cont'd)

521	120	4.34	4.42	-.27	.08	1.22	1.6	1.17	1.2	1.00	.75	.70	27.9	26.0	14 14	
522	120	4.35	4.43	-.27	.08	1.24	1.8	1.29	2.1	.95	.61	.70	27.0	26.0	12 12	
522	120	4.35	4.43	-.27	.08	1.19	1.4	1.18	1.3	.88	.79	.70	25.9	26.0	48 48	
523	120	4.36	4.44	-.28	.08	1.10	.8	1.07	.5	1.01	.71	.70	26.1	26.0	78 78	
528	120	4.40	4.50	-.31	.08	.99	.0	1.09	.6	.97	.68	.70	25.8	26.1	34 34	
529	120	4.41	4.51	-.32	.08	.67	-2.9	.65	-3.0	1.58	.71	.70	30.2	26.1	29 29	
530	120	4.42	4.52	-.32	.08	.61	-3.5	.63	-3.2	1.32	.76	.70	28.7	26.1	79 79	
532	120	4.43	4.54	-.34	.08	1.26	1.9	1.36	2.4	.98	.66	.70	28.2	26.2	77 77	
533	120	4.44	4.55	-.34	.08	.81	-1.5	.85	-1.1	1.30	.73	.70	28.5	26.2	56 56	
536	120	4.47	4.58	-.36	.08	.92	-.5	.92	-.5	1.28	.78	.70	28.2	26.2	71 71	
537	120	4.47	4.59	-.37	.08	1.10	.7	1.17	1.2	.79	.57	.70	23.6	26.2	62 62	
540	120	4.50	4.63	-.39	.08	.57	-4.0	.66	-2.9	1.11	.68	.70	26.5	26.2	3 3	
540	120	4.50	4.63	-.39	.08	.88	-.9	.90	-.7	1.28	.85	.70	29.0	26.2	75 75	
541	120	4.51	4.64	-.39	.08	1.03	.2	1.10	.7	1.16	.63	.70	27.5	26.2	47 47	
541	120	4.51	4.64	-.39	.08	1.64	4.2	1.70	4.4	.55	.57	.70	24.1	26.2	63 63	
543	120	4.53	4.66	-.41	.08	.98	-.1	1.00	.0	1.05	.75	.70	28.3	26.2	80 80	
544	120	4.53	4.67	-.41	.08	1.94	5.8	1.95	5.7	.22	.30	.70	20.4	26.3	15 15	
544	120	4.53	4.67	-.41	.08	.95	-.3	1.13	1.0	.91	.40	.70	23.9	26.3	25 25	
545	120	4.54	4.68	-.42	.08	1.23	1.7	1.30	2.1	.67	.65	.70	24.9	26.3	57 57	
546	120	4.55	4.69	-.43	.08	.60	-3.6	.60	-3.5	1.42	.76	.70	28.8	26.3	38 38	
553	120	4.61	4.76	-.47	.08	1.18	1.3	1.12	.9	.88	.74	.70	25.4	26.3	74 74	
557	120	4.64	4.80	-.50	.08	.71	-2.4	.73	-2.2	1.29	.78	.69	28.6	26.3	1 1	
557	120	4.64	4.80	-.50	.08	1.05	.4	1.04	.3	.94	.82	.69	25.8	26.3	9 9	
557	120	4.64	4.80	-.50	.08	.95	-.3	1.10	.7	.85	.72	.69	26.4	26.3	32 32	
558	120	4.65	4.81	-.50	.08	.91	-.6	.96	-.2	1.11	.79	.69	27.6	26.3	20 20	
558	120	4.65	4.81	-.50	.08	.56	-4.0	.52	-4.4	1.49	.81	.69	29.4	26.3	51 51	
558	120	4.65	4.81	-.50	.08	.59	-3.7	.59	-3.6	1.46	.84	.69	30.0	26.3	69 69	
559	120	4.66	4.82	-.51	.08	1.10	.8	1.16	1.1	.94	.69	.69	27.7	26.3	46 46	
561	120	4.68	4.84	-.52	.08	1.55	3.7	1.47	3.1	.59	.71	.69	25.1	26.3	16 16	
561	120	4.68	4.84	-.52	.08	.57	-3.9	.56	-3.9	1.46	.78	.69	29.1	26.3	55 55	
567	120	4.72	4.90	-.56	.08	.53	-4.4	.58	-3.7	1.37	.71	.69	28.3	26.2	36 36	
569	120	4.74	4.92	-.58	.08	1.42	2.9	1.49	3.2	.73	.55	.69	25.8	26.2	40 40	
569	120	4.74	4.92	-.58	.08	.72	-2.3	.75	-2.0	1.30	.80	.69	28.2	26.2	45 45	
570	120	4.75	4.93	-.58	.08	1.30	2.1	1.33	2.3	.54	.73	.69	23.2	26.2	67 67	
575	120	4.79	4.98	-.62	.08	1.02	.1	1.01	.1	.99	.80	.69	27.3	26.1	23 23	
576	120	4.80	4.99	-.62	.08	1.04	.3	1.17	1.2	1.01	.72	.69	28.0	26.1	52 52	
579	120	4.82	5.02	-.64	.08	1.64	4.1	1.51	3.3	.58	.69	.69	23.9	26.1	76 76	
586	120	4.88	5.09	-.69	.08	.91	-.7	.92	-.5	1.01	.76	.68	25.8	26.0	17 17	
587	120	4.89	5.09	-.70	.08	1.44	3.0	1.36	2.4	.67	.75	.68	24.5	26.0	83 83	
591	120	4.93	5.13	-.73	.08	1.03	.2	1.17	1.2	.96	.70	.68	27.7	25.9	19 19	
592	120	4.93	5.14	-.73	.08	.76	-1.9	.84	-1.1	1.18	.73	.68	27.9	25.9	27 27	
592	120	4.93	5.14	-.73	.08	.96	-.3	.92	-.6	1.12	.70	.68	27.0	25.9	64 64	
594	120	4.95	5.16	-.75	.08	1.27	1.9	1.40	2.6	.74	.62	.68	24.2	25.8	6 6	
606	120	5.05	5.27	-.83	.08	1.27	1.9	1.47	3.0	.54	.58	.67	24.1	25.5	30 30	
617	120	5.14	5.36	-.91	.09	.77	-1.8	.75	-1.9	1.15	.75	.66	26.4	25.1	54 54	
633	120	5.28	5.50	-1.03	.09	1.09	.7	.95	-.2	1.08	.75	.65	26.0	24.5	53 53	
649	120	5.41	5.62	-1.15	.09	.64	-3.0	.53	-4.1	1.48	.71	.64	27.2	23.8	72 72	
655	120	5.46	5.67	-1.20	.09	.77	-1.7	.91	-.6	1.08	.68	.63	24.2	23.5	60 60	
660	120	5.50	5.71	-1.24	.09	1.44	2.8	1.33	2.1	.78	.57	.63	23.8	23.2	68 68	
661	120	5.51	5.71	-1.25	.09	.65	-2.9	.64	-2.9	1.31	.71	.63	24.7	23.1	4 4	
668	120	5.57	5.77	-1.31	.09	1.01	.1	.93	-.4	1.23	.70	.62	24.9	22.7	82 82	
705	120	5.88	6.04	-1.65	.10	1.05	.3	.97	-.1	1.12	.58	.58	21.8	20.3	50 50	
536.2	120.0	4.47	4.56	-.38	.08	1.00	-.2	1.04	.1		.69				Mean (Count: 83)	
59.0	.0	.49	.59	.39	.00	.31	2.4	.33	2.4		.10				S.D. (Population)	
59.3	.0	.49	.59	.40	.00	.31	2.4	.33	2.4		.10				S.D. (Sample)	

Model, Populn: RMSE .08 Adj (True) S.D. .39 Separation 4.75 Strata 6.66 Reliability (not inter-rater) .96
Model, Sample: RMSE .08 Adj (True) S.D. .39 Separation 4.78 Strata 6.70 Reliability (not inter-rater) .96
Model, Fixed (all same) chi-squared: 1770.0 d.f.: 82 significance (probability): .00
Model, Random (normal) chi-squared: 78.3 d.f.: 81 significance (probability): .56
Inter-Rater agreement opportunities: 408360 Exact agreements: 104870 = 25.7% Expected: 103317.0 = 25.3%

APPENDIX J: INTERSECTIONS OF NONVERBAL BEHAVIOR

Table J.1

Percentages of Behavior Across Categories of Affect

	Anxiety	Attentive	Attitude	Competence	Confidence	Desire to Communicate	Engagement	Expressiveness	Happiness	Humor	Interactiveness	Warmth	Raw Total
Body Language (General)	32	5	5	5	26	0	11	5	5	0	0	5	19
Eyebrows	14	14	5	0	10	0	14	19	5	0	14	5	19
Face (General)	19	8	7	6	13	3	5	19	8	0	4	7	108
Averted gaze	26	13	4	4	21	0	17	2	2	0	6	4	47
Blinking	0	0	0	0	0	0	0	0	0	0	0	0	0
Eyes grow wide	33	17	0	0	0	0	0	17	17	0	0	17	6
Mutual gaze	10	17	5	6	10	3	25	11	3	0	6	3	63
Shifting gaze	35	4	0	4	20	4	11	12	1	1	7	1	82
Staring	20	13	7	0	7	0	7	13	13	0	7	13	15
Unfocused gaze	0	0	0	100	0	0	0	0	0	0	0	0	1
Lack of hand movement	16	5	11	11	11	0	5	11	11	0	21	0	19
Random movement	50	0	0	0	17	0	0	17	0	0	17	0	6
Representational	0	0	0	6	0	6	13	31	19	0	13	13	16
Self-adaptor	55	5	0	0	14	0	14	5	5	0	0	5	22
Head turn	21	7	7	0	7	14	21	14	0	0	0	7	14
Nodding	13	18	7	2	13	0	18	9	4	0	11	4	45
Laughing	21	4	11	3	6	4	1	14	11	6	3	15	71
Frowning	20	20	0	20	20	0	20	0	0	0	0	0	5
Lack of smile	3	3	22	13	6	6	0	3	25	0	3	16	32
Lip Movements	33	0	17	0	17	0	8	17	8	0	0	0	12
Mouth barely open	33	0	0	0	33	0	0	33	0	0	0	0	3
Nervous smile	33	0	6	6	22	6	0	0	17	0	6	6	18
Smile	13	6	13	3	10	2	6	10	20	1	3	11	174
Swallowing	31	0	15	0	0	0	8	8	8	8	0	23	13
Audible breathing	100	0	0	0	0	0	0	0	0	0	0	0	2
Backchannel	0	20	20	0	0	0	20	0	20	0	0	20	5
Filled pauses	34	9	0	11	23	0	14	3	0	0	3	3	35
Tone-Prosody	13	4	6	4	9	4	6	21	13	6	2	13	53
Volume	0	33	0	0	33	0	0	0	0	0	0	33	3
Adjusting posture	38	0	13	0	25	0	0	0	0	13	0	13	8
Leaning back-Slouching	33	0	0	0	33	0	33	0	0	0	0	0	9
Leaning forward	6	24	0	0	18	6	29	6	0	0	6	6	17
Moving around	25	0	5	0	20	5	15	15	0	5	5	5	20
Rigid/Straight	21	5	5	0	16	0	11	21	11	0	5	5	19
Rocking-Shaking	65	5	0	0	10	0	10	5	5	0	0	0	20
Shoulders	50	0	17	0	17	0	0	17	0	0	0	0	6

Table J.2

Percentages of Behavior Within Categories of Affect

	Anxiety	Attentive	Attitude	Competence	Confidence	Desire to Communicate	Engagement	Expressiveness	Happiness	Humor	Interactiveness	Warmth
Body Language (General)	3	1	1	2	4	0	2	1	1	0	0	1
Eyebrows	1	4	1	0	2	0	3	3	1	0	6	1
Face (General)	10	12	11	17	11	13	5	17	10	0	8	10
Averted gaze	6	8	3	5	8	0	8	1	1	0	6	3
Blinking	0	0	0	0	0	0	0	0	0	0	0	0
Eyes grow wide	1	1	0	0	0	0	0	1	1	0	0	1
Mutual gaze	3	14	4	10	5	8	16	6	2	0	8	3
Shifting gaze	13	4	0	7	12	13	9	9	1	8	13	1
Staring	1	3	1	0	1	0	1	2	2	0	2	3
Unfocused gaze	0	0	0	2	0	0	0	0	0	0	0	0
Lack of hand movement	1	1	3	5	2	0	1	2	2	0	8	0
Random movement	1	0	0	0	1	0	0	1	0	0	2	0
Representational	0	0	0	2	0	4	2	4	3	0	4	3
Self-adaptor	6	1	0	0	2	0	3	1	1	0	0	1
Head turn	1	1	1	0	1	8	3	2	0	0	0	1
Nodding	3	11	4	2	5	0	8	3	2	0	10	3
Laughing	7	4	11	5	3	13	1	9	9	33	4	14
Frowning	0	1	0	2	1	0	1	0	0	0	0	0
Lack of smile	0	1	10	10	2	8	0	1	9	0	2	6
Lip Movements	2	0	3	0	2	0	1	2	1	0	0	0
Mouth barely open	0	0	0	0	1	0	0	1	0	0	0	0
Nervous smile	3	0	1	2	3	4	0	0	3	0	2	1
Smile	11	14	32	14	13	13	11	16	38	8	13	26
Swallowing	2	0	3	0	0	0	1	1	1	8	0	4
Audible breathing	1	0	0	0	0	0	0	0	0	0	0	0
Backchannel	0	1	1	0	0	0	1	0	1	0	0	1
Filled pauses	6	4	0	10	6	0	5	1	0	0	2	1
Tone-Prosody	3	3	4	5	4	8	3	9	8	25	2	9
Volume	0	1	0	0	1	0	0	0	0	0	0	1
Adjusting posture	1	0	1	0	2	0	0	0	0	8	0	1
Leaning back-Slouching	1	0	0	0	2	0	3	0	0	0	0	0
Leaning forward	0	5	0	0	2	4	5	1	0	0	2	1
Moving around	2	0	1	0	3	4	3	3	0	8	2	1
Rigid/Straight	2	1	1	0	2	0	2	3	2	0	2	1
Rocking-Shaking	6	1	0	0	2	0	2	1	1	0	0	0
Shoulders	1	0	1	0	1	0	0	1	0	0	0	0
Raw Total	217	76	73	42	129	24	100	116	92	12	48	78

Table J.3

Percentages of Behavior Across Categories of Language

	Comprehensibility	Comprehension	Fluency	Grammar	Pronunciation	Vocabulary	Raw Total
Body Language (General)	25	0	25	0	25	25	4
Eyebrows	0	50	0	25	25	0	4
Face (General)	10	46	22	5	5	10	41
Averted gaze	22	28	39	6	0	22	18
Blinking	0	0	0	0	0	0	0
Eyes grow wide	0	100	0	0	0	0	2
Mutual gaze	12	31	19	15	8	12	26
Shifting gaze	13	21	40	4	6	13	47
Staring	0	50	50	0	0	0	2
Unfocused gaze	0	0	0	0	0	0	0
Lack of hand movement	17	17	33	17	0	17	6
Random movement	33	0	67	0	0	33	3
Representational	0	100	0	0	0	0	1
Self-adaptor	33	0	33	0	0	33	6
Head turn	22	22	33	11	11	22	9
Nodding	0	73	9	0	9	0	11
Laughing	15	15	40	10	5	15	20
Frowning	0	100	0	0	0	0	1
Lack of smile	0	0	25	25	50	0	4
Lip Movements	0	0	43	14	14	0	7
Mouth barely open	25	0	0	0	50	25	4
Nervous smile	0	25	0	0	25	0	4
Smile	9	34	14	11	9	9	44
Swallowing	0	0	50	0	0	0	2
Audible breathing	0	0	100	0	0	0	1
Backchannel	0	0	0	0	0	0	0
Filled pauses	11	11	56	14	0	11	57
Tone-Prosody	7	11	22	7	44	7	27
Volume	0	0	0	0	100	0	1
Adjusting posture	0	50	50	0	0	0	4
Leaning back-Slouching	0	0	0	0	0	0	0
Leaning forward	0	50	33	0	0	0	6
Moving around	36	0	9	0	36	36	11
Rigid/Straight	0	0	0	0	100	0	1
Rocking-Shaking	13	13	50	13	0	13	8
Shoulders	25	25	0	0	25	25	4

Table J.4

Percentages of Behavior Within Categories of Language

	Comprehensibility	Comprehension	Fluency	Grammar	Pronunciation	Vocabulary
Body Language (General)	25	0	25	0	25	25
Eyebrows	0	50	0	25	25	0
Face (General)	10	46	22	5	5	10
Averted gaze	22	28	39	6	0	22
Blinking	0	0	0	0	0	0
Eyes grow wide	0	100	0	0	0	0
Mutual gaze	12	31	19	15	8	12
Shifting gaze	13	21	40	4	6	13
Staring	0	50	50	0	0	0
Unfocused gaze	0	0	0	0	0	0
Lack of hand movement	17	17	33	17	0	17
Random movement	33	0	67	0	0	33
Representational	0	100	0	0	0	0
Self-adaptor	33	0	33	0	0	33
Head turn	22	22	33	11	11	22
Nodding	0	73	9	0	9	0
Laughing	15	15	40	10	5	15
Frowning	0	100	0	0	0	0
Lack of smile	0	0	25	25	50	0
Lip Movements	0	0	43	14	14	0
Mouth barely open	25	0	0	0	50	25
Nervous smile	0	25	0	0	25	0
Smile	9	34	14	11	9	9
Swallowing	0	0	50	0	0	0
Audible breathing	0	0	100	0	0	0
Backchannel	0	0	0	0	0	0
Filled pauses	11	11	56	14	0	11
Tone-Prosody	7	11	22	7	44	7
Volume	0	0	0	0	100	0
Adjusting posture	0	50	50	0	0	0
Leaning back-Slouching	0	0	0	0	0	0
Leaning forward	0	50	33	0	0	0
Moving around	36	0	9	0	36	36
Rigid/Straight	0	0	0	0	100	0
Rocking-Shaking	13	13	50	13	0	13
Shoulders	25	25	0	0	25	25
Raw Total	46	95	119	32	41	53