Studies in Second Language Acquisition (2025), 1–26
doi:10.1017/S0272263125000063

RESEARCH ARTICLE

Affect as a component of second language speech perception

John Dylan Burton[1] and Paula Winke[2]

[1]Applied Linguistics and ESL, Georgia State University, Atlanta, GA, USA and [2]Second Language Studies, Michigan State University, East Lansing, MI, USA
Corresponding author: John Dylan Burton; Email: jdburton@gsu.edu

(Received 20 May 2024; Revised 08 October 2024; Accepted 24 December 2024)

Abstract

Growing evidence suggests that ratings of second language (L2) speech may be influenced by perceptions of speakers' affective states, yet the size and direction of these effects remain underexplored. To investigate these effects, 83 raters evaluated 30 speech samples using 7-point scales of four language features and ten affective states. The speech samples were 2-min videorecordings from a high-stakes speaking test. An exploratory factor analysis reduced the affect scores to three factors: assuredness, involvement, and positivity. Regression models indicated that affect variables predicted spoken language feature ratings, explaining 18–27% of the variance in scores. Assuredness and involvement corresponded with all language features, while positivity only predicted comprehensibility scores. These findings suggest that listeners' perceptions of speakers' affective states intertwine with their spoken language ratings to form a visual component of second-language communication. The study has implications for models of L2 speech, language pedagogy, and assessment practice.

Keywords: affect; language assessment; proficiency; speech perception

Introduction

Second language (L2) speakers are often evaluated by the world around them—implicitly or explicitly—on their capacity to communicate in social situations. Comprehensibility, speech fluency, or the accuracy of grammar or vocabulary may influence how proficient someone is perceived in their L2. These perceptions can then alter an individual's prospects of succeeding at various low and high-stakes real-world tasks that require language. For example, if someone's language ability is perceived as "inadequate," they may face stigma or discrimination by not receiving raises, being passed over in interviews, or even being fired from their place of employment (Gluszek & Dovidio, 2010; Kang & Yaw, 2024). Language ability is important in L2 research settings as well, as learners are often grouped by proficiency profiles in research on a

myriad of acquisitional processes. It is often assumed that linguistic features of speech drive perceptions of language ability, and there is a vast body of research supporting this claim. However, it is possible—indeed, even intuitive—that non-linguistic factors such as emotion and body language influence these perceptions as well, yet research in this area is currently scant.

In many settings, impressions about individuals' cognitive and psychological states and traits are influenced by nonverbal behaviour and its corresponding affective interpretation, such as how confident, anxious, happy, or engaged a person appears. A confident and engaged speaker may be perceived as able to handle a communicative situation more adeptly, for example, than a speaker who is relatively reticent and anxious, despite similarities in actual language produced (Jenkins & Parra, 2003). Hymes (1972, p. 283) theorized that affective dispositions (along with cognitive and volitive factors) are elements of "ability for use" that moderate how individuals deploy linguistic knowledge in communication. These factors could then partially determine an individual's assessed level of communicative competence in a particular scenario or sociolinguistic context. In nearly all situations, listeners detect affect displays, even if they are unaware of doing so (Bargh & Chartrand, 1999). This combined visual information has even been found to lead to score differences in language tests, where audiovisual speech samples are perceived as exhibiting higher comprehensibility or overall ability than the same individuals in audio-only samples (Carey & Szocs, 2024; Nakatsuhara et al., 2021).

If perceptions of affect form a visual component of listeners' mental model of L2 speech, these relationships should be documented, as decisions based on these perceived abilities may impact social outcomes and interpretations of research findings. This study sets out to measure speakers' perceptions of both affect and spoken language ability to determine the extent to which these non-linguistic and linguistic elements are related. Understanding these relationships may have important practical and theoretical implications for how people teach, assess, and model second language speech.

Background

Affect generally refers to the subjective experience of internal feelings, emotions, moods, dispositions, and temperaments, often visible through behaviour (Frijda, 1994). Individuals show affect through displays—that is, intentional or unintentional behaviours that convey an individual's orientations or reactions to stimuli and other people. Although communicated verbally through word choice or prosody, affect is primarily conveyed through a speaker's facial expressions (Kappas et al., 2013). A speaker may display a feeling, orientation, or stance at one moment as a *state*, or they may be disposed to react to particular experiences in a similar way over time as a *trait* element of their personality. Affect was historically thought to originate in individual cognition, whereby appraisals of environmental stimuli activate emotional responses (Arnold, 1960). Others have argued that due to the lack of one-to-one physiological correlates, affect may arise within socially distributed processes amongst individuals in interpersonal interactions (Parkinson, 1996). Furthermore, interpretations of affect may differ depending on the cultural background of speakers and listeners (Uchida et al., 2009) and may even be *felt* differently across cultures (Mesquita, 2022). These varying issues highlight methodological challenges in the measurement of affect;

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

the social context should be carefully documented, and cultural variables should be controlled if possible.

Affect displays present distinct challenges for L2 speakers in cross-cultural settings because, to communicate effectively, they must "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'" (Pavlenko, 2014, p. 247). Furthermore, the presence of internal affective responses may have a facilitating or limiting effect on the language they produce, 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). Affect may then drive differential performance outcomes in learners based on their ever-changing reactions to social stimuli.

Indeed, trait-like affective factors have received much attention in the second language acquisition (SLA) literature because they have been found to relate to various pedagogical outcomes. Higher measures of anxiety (e.g., foreign language anxiety, test anxiety), for example, have been found to correspond with lower levels of language proficiency and course achievement (Botes et al., 2020; MacIntyre et al., 1997; Teimouri et al., 2019). In contrast, confidence (especially L2 *self*-confidence) has a positive relationship with language knowledge measures and performance outcomes (Ahammer et al., 2019; Clément, 1986; Noels & Clément, 1996; Stankov et al., 2012), perhaps given its close relationship with other individual differences such as motivation and willingness to communicate (MacIntyre et al., 1998). Positive emotions (e.g., enjoyment, happiness) may also drive or correspond with L2 acquisitional processes by "broadening a person's perspective and opening the individual to absorb the language" (MacIntyre & Gregersen, 2012, p. 193), with some evidence that these correspond with achievement gains as well (Botes et al., 2020; Dewaele & Li, 2022; Li et al., 2020). Nonetheless, the relationship between at least some of these measures and achievement or proficiency outcomes may be reciprocal or bidirectional, with growth in language skills leading to, for example, greater self-confidence in one's ability to communicate (Edwards & Roger, 2015; Li et al., 2020). In many of these studies, these affective traits were measured using student self-report surveys rather than external observations, which offer limited evidence about how listeners conceptualize spoken language ability considering dynamic differences in affect.

Although these affective traits lend important insight into longer-term outcomes in SLA such as course achievement, perceived affective states may impact outcomes as performances unfold in brief encounters, such as conversations or interviews. Engagement, for example, broadly defined as an individual's level of interest and evidence of participation in an event, is often regarded as an orientation composed of interacting cognitive, social, behavioural, and affective dimensions (Philp & Duchesne, 2016). Engagement, in particular social engagement, may also be critical to how successful individuals are in interactive tasks, leading to enhanced performance outcomes (Storch, 2008). Being perceived as engaged, as well as displaying confidence, attentiveness, interactiveness, and low anxiety, may factor into positive impressions of communicative effectiveness in spoken assessment settings (Ducasse & Brown, 2009; May, 2011; Sato & McNamara, 2019). Other affective phenomena, such as displaying warmth (e.g., friendliness, empathy) or competence, have been found to correspond with performance outcomes outside of SLA research in organizational settings (Cuddy et al., 2011).

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

Few empirical studies to date have measured the relationships between perceived state affect and language-related outcomes. Nagle et al. (2022) considered subjectively perceived measures of anxiety and collaborativeness (a measure of social engagement and interaction) on scores of L2 comprehensibility. In a dataset of short dyadic interactions, speakers and their interlocutors each repeatedly measured their partner's perceived affective states and comprehensibility. The authors found that high collaborativeness (a measure of social engagement) and low anxiety explained roughly 60% of the variance in comprehensibility scores, with small differences depending on the type of speaking task used. The authors hypothesized that high anxiety related to lower comprehensibility due to the visual cues anxious speakers display (e.g., lack of expressiveness, gaze aversion), which make speech processing more effortful for the listeners. In another study, Chong and Aryadoust (2023) investigated the relationship between automated measurements of seven basic emotions (happiness, sadness, anger, surprise, fear, disgust, and a neutral state) and language proficiency outcomes provided by four human raters using TOEFL integrated speaking rubrics. In this study, the variance in test scores attributable to emotions ranged from 8–34%, with the authors concluding that "only some part of the observed variance in test scores can or should be associated with the emotions of participants when they are using academic language" (p. 7431). However, the researchers did not interpret which emotions were associated more consistently with language-related outcomes.

The current study

Although the literature has shown a possible link between perceived affect and language-related outcomes, affect is often treated as a trait variable in individual differences studies, often measured through self-report data (Botes et al., 2020; Clément, 1986; Dewaele & Li, 2022; Li et al., 2020; MacIntyre et al., 1997; Noels & Clément, 1996; Teimouri et al., 2019). Studies that have operationalized affect as a state variable have often relied on verbal reports focusing on communication as a whole (Sato & McNamara, 2019) or interactional competence more narrowly (Ducasse & Brown, 2009; May, 2011). While these studies noted the affective component of L2 speech, little is known about the size of its relationship. What empirical research exists has investigated outcomes of comprehensibility (Nagle et al., 2022) and integrated skill ratings of spoken language proficiency (Chong & Aryadoust, 2023). A range of other components of language proficiency, such as fluency, grammar, and lexical range and accuracy, have yet to be considered. It is important to consider a broader range of variables given that affect may be related to some language skills more than others (Dewaele & Li, 2022).

Given the relatively understudied role affect may play in ratings of L2 speech, the research question (RQ) that guided this study was the following:

> RQ: What are the relationships, if any, between ratings of L2 speech and ratings of affective phenomena?

Based on the literature reviewed, we hypothesized that affective perceptions such as confidence, engagement, and low anxiety would correspond with higher ratings in language outcomes broadly. Overall, though, our investigation was exploratory as we had few expectations regarding the direction and size of the effects.

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

Method

Participants

After obtaining ethics clearance through our university's institutional review board (ID number: STUDY00006268), we invited 100 participants to take part in this study. These participants were individuals with limited experience working in language-related settings; that is to say, linguistic laypeople (Sato & McNamara, 2019). The use of this population aligns with research in SLA that has investigated relationships with language ratings using novice raters (e.g., Isaacs & Trofimovich, 2012). These "naïve" listeners, rather than trained language educators or researchers, were chosen to provide observations that would better reflect how individuals in society incorporate affect into their language-related judgements. Because the cultural and linguistic backgrounds of speakers play a role in how facial nonverbal behaviour encodes affect and is decoded by listeners (Matsumoto & Hwang, 2016), participants' backgrounds were controlled to reduce this source of variance: all participants were first language (L1) English, USA-born undergraduates at a large public university. The mean age of the listener-raters was 20.92 years ($SD = 1.48$), with a roughly balanced distribution of gender (52% female, 41% male; 6% indicated a gender identification other than male/female or preferred not to report). Approximately one third of the participants indicated knowledge of an L2 (38%), and their indicated areas of study were diverse, as the invitation to participate in the study was sent to all undergraduates on campus.

The sample sizes for both the participant raters and the speech samples, described in the next section, were determined using a power analysis and a reading of the literature (Hox et al., 2018). Hox et al. (2018) suggested that a greater number of second-level grouping variables (raters in this study) would generally provide more power than a greater number of first-level cases (speech samples). A sample of at least 80 raters and 30 speech samples were determined to provide power at .95 to detect small to medium regression coefficients. We overrecruited participants because we anticipated that some novice raters would exhibit less reliable rating patterns. This turned out to be the case, as we found that 16 of the 100 participants exhibited undesirable rating qualities. Some participants showed outlying rating patterns measured using multivariate outlier analysis, and others showed misfit with a many-facet Rasch measurement model. We removed these 16 raters plus one rater who experienced technical problems. This served to optimize the quality of the dataset, leaving a final number of 83 raters for the current study. We report the data cleaning procedures in detail in Supplement 1 in the Open Science Framework (Burton & Winke, 2025).

Speech samples

We borrowed 30 speech samples recorded in a high-stakes, oral proficiency interview context from a test provider (International English Language Testing System [IELTS]; IELTS, n.d.) to be used as the basis for the affect and spoken language ability ratings. We signed non-disclosure agreements with IELTS to protect the dataset and the privacy of the test takers, and all raters signed non-disclosure agreements indicating that they would not remove or report on the samples they watched. The test takers in this dataset had indicated consent for their data to be used in research, but because of the intellectual property of the testing data, we were unable to share video or audio from the test sessions with readers. High-stakes test recordings were optimal as many contextual elements were controlled: the recordings were conducted on a standardized laptop, the noise was controlled in the environment, the IELTS-employed examiners

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

who conducted the interviews were trained, and the language of the test content was validated for the ability levels of the test takers and test purposes.

The recordings were taken from the same section of the speaking test (Part 3) for all samples. The section was a semi-scripted conversation between a trained IELTS examiner and the test taker on abstract issues and ideas (IELTS, n.d.). Although the interview topics and examiners varied, the samples contained a similar number of opportunities for the test takers to speak and clarify answers across the segments. We selected segments of approximately 2 mins from each test taker (M = 2 min, 11 s; SD = 14 s) from the beginning of this part of the test. The length of these segments varied slightly because we sought to trim the samples as close to the 2-min mark as possible when the test taker had reached a natural conclusion to their turn. We chose relatively short (2-min) samples from the longer test as the basis of rating so that participants would make quick, intuitive impressions rather than impressions based on a wider range of evidence. Using short samples also allowed us to keep the total experiment participation time for the volunteer raters at around 2 hours.

Table 1 displays information about the speech samples. The individuals in the 30 samples (labelled S01–S30) were all Chinese L2 English speakers. Their sample label indicated their test score ranking in comparison with other test takers. For example, sample S01 had the lowest proficiency test score on IELTS (3.5, approximately A2 on the Common European Framework of Reference [CEFR]; Council of Europe,

Table 1. Speech samples

Sample	Gender	Test Score	Duration (mm:ss)
S01	M	3.5	02:23
S02	M	4	02:31
S03	M	4	02:19
S04	F	4	01:43
S05	F	4	02:09
S06	F	4	02:18
S07	F	4	02:11
S08	M	4.5	01:53
S09	F	5	02:00
S10	F	5	02:31
S11	F	5	01:46
S12	F	5	01:49
S13	M	5	02:13
S14	F	5	02:02
S15	F	5.5	02:22
S16	M	5.5	02:19
S17	F	5.5	01:51
S18	F	5.5	02:18
S19	F	5.5	02:05
S20	F	5.5	01:57
S21	F	5.5	02:23
S22	M	5.5	02:13
S23	F	6	02:20
S24	F	6	02:43
S25	F	6	02:14
S26	F	6	02:13
S27	F	6.5	02:14
S28	F	6.5	01:53
S29	F	6.5	02:22
S30	F	6.5	02:15

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

2020) and was thus labelled as 01. On the other hand, sample S30 tied for the highest score (6.5, approximately a high B2/low C1 on the CEFR), and was listed last (IELTS scores range from 0 to 9). The test takers appeared to be of a similar age (approximately college age), and the distribution of gender appeared to skew female (23 females, 7 males), but these demographic variables were not provided with the dataset. Proficiency scores that accompanied the dataset showed that test takers were evenly distributed across multiple ability levels. The last column in Table 1 shows the length of each of the speech samples. More information about how the dataset was compiled and trimmed is available in Supplement 2 (Burton & Winke, 2025).

Rating scales

The rater participants in this study assigned subjective ratings on affect and spoken language ability using a set of 14 *semantic differential* scales. Semantic differential scales are simple, often one-word adjectives or nouns that are paired with a contrasting term set on two ends of the same scale, similar to a Likert scale (e.g., good/bad, interesting/uninteresting) (Ploder & Eder, 2015). These scales may have multiple points between terms for raters to indicate both the directionality and strength of an association with a term. Semantic differentials allow participants to make quick, intuitive decisions on their perceptions of stimuli that do not require rater training, unlike fully developed rating rubrics/scales (Snider & Osgood, 1969). Semantic differentials also allow participants to bring their understanding and interpretation of phenomena to a rating event rather than restricting these interpretations through training or scale wording. This subjective, relatively open approach was desirable for this study to capture more generalizable impressions of speech and affect.

We constructed a set of scales with categories describing spoken language performance and interpersonal perceptions of affective states. The language features were fluency, vocabulary, grammar, and comprehensibility, roughly corresponding to categories of language proficiency frequently of interest in the SLA literature. Comprehensibility, rather than pronunciation, was chosen as a category to align with ongoing research in this area (e.g., Isaacs & Trofimovich, 2012; Nagle et al., 2022). Participants may have had more of an intuitive understanding of what is comprehensible rather than constituent elements of pronunciation (e.g., phonemic control, prosody), which may have led to an overt focus on accent. Even though we determined through piloting that these four categories were relatively straightforward for participants to understand, we provided brief definitions of these four categories to reduce ambiguity (e.g., broad/narrow definitions of fluency; Lennon, 1990). These brief definitions are listed in Table 2. More information about the piloting process and how the scales were developed is available in Burton (2023).

We chose 10 categories of affect for participants to rate based on frequently discussed affective state perceptions discussed in the literature on SLA, language

Table 2. Definitions of scale categories

Category	Definition
Fluency	Rate of speech, breakdowns, and repair
Vocabulary	Range, accuracy, and complexity of words
Grammar	Range, accuracy, and complexity of grammar
Comprehensibility	How difficult/easy it is to understand the person

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

testing, and psychology. These categories were: engagement, anxiety, confidence, warmth, attentiveness, expressiveness, happiness, competence, interactiveness, and attitude. Of particular interest were those states that have been discussed in relation to language proficiency or achievement, such as engagement (Ducasse & Brown, 2009; Jenkins & Parra, 2003; May, 2011; Nakatsuhara et al., 2021; Philp & Duchesne, 2016; Sato & McNamara, 2019), anxiety (Botes et al., 2020; McIntyre et al., 1997; Sato & McNamara, 2019; Teimouri et al., 2019), and confidence (Clément, 1986; Ducasse & Brown, 2009; Noels & Clément, 1996; May, 2011). We were also interested in how socio-affective perceptions of speakers might relate to language judgements, and for this reason, we included attentiveness (Ducasse & Brown, 2009; May, 2011) and interactiveness (relating to interactional competence; Galaczi & Taylor, 2018; and also, willingness to communicate; MacIntyre et al., 1999). In addition, we included a measure of expressiveness, as this has been frequently mentioned as a positive element in relation to proficiency judgements when referring to overall nonverbal behaviour (e.g., Jenkins & Parra, 2003; Neu, 1990), as well as happiness and attitude, relating to measures of enjoyment and positive psychology in the SLA literature (Botes et al., 2020; Dewaele & Li, 2022; Li et al., 2020; MacIntyre & Gregersen, 2012; MacIntyre et al., 2019). Finally, warmth and competence were chosen based on findings in psychology relating these perceptions to success in organizational settings (Cuddy et al., 2011). As opposed to the language scales, definitions were not provided for the affect-related adjectives, as raters were expected to bring their own interpretations of these variables to the study.

Each scale category was presented with its adjective and the related antonym on a 7-point scale, as shown in Figure 1. We used a 7-point scale to enhance measurement precision over scales with fewer categories (Simms et al., 2019). This also allowed a midpoint for cases where judgements were ambiguous. The polarity of the adjectives was reversed for half of the scales so that positive or negative associations were mixed on each side of the 7-point scale. This was to reduce survey acquiescence bias (e.g.,

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*

Figure 1. Rating scales.

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

straightlining), and to encourage participants to read each line carefully. The language scales were presented earliest and in the same order to establish the primacy of rating language. The affect scales were presented below the language scales in a random order for each speech sample rating. The random order was to prevent primacy of any of the affect categories, while also encouraging raters to pay close attention to the category they were rating each time. To determine the feasibility of the instruments and the design of the rating study, the scales were piloted with 25 participants (not included in this rater group) with a separate set of 10 speech samples collected in a prior study. A many-facet Rasch analysis of the pilot data indicated that each scale functioned well, with scale units ordered as intended with no misfit. The pilot participants indicated that the scales were straightforward to use when rating, and the number of scales was not an issue given the quick nature of the rating process. Specific details regarding the pilot data analysis are available in Burton (2023).

Procedure

The videos and scales were incorporated into an online rating platform built in Qualtrics. An example of the system is presented in Supplement 3 (Burton & Winke, 2025). Raters were first introduced to the study, signed a consent form and a non-disclosure agreement, and were then allowed to review the scale categories and their meanings. Although language terms were defined briefly, affect terms were not, as semantic differentials generally allow users to bring their internal definitions of terms (Ploder & Eder, 2015). Participants were required to rate two practice videos using the scales to calibrate their orientations to spoken language proficiency. One video was of a highly proficient speaker, and the other was of a much less proficient speaker. After rating, participants were not provided with feedback on their scores, but rather with a general description of the language performances (not including affect), as it was desirable for participants to focus first and foremost on spoken language ability. The descriptions were framed in terms of the strength of language and communicative effectiveness. After completing the practice section, raters began the main study.

The study took place on two different days with a 24-hr gap between them. Each day featured 15 randomized speech samples for each participant. The rating was spread out to minimize fatigue, as each rating day took roughly 1 hr to complete (Day 1: $M = 61$ min, $SD = 18$ min; Day 2 = 62 min, $SD = 20$ min). Participants conducted the ratings remotely, though they were instructed to choose a location that was quiet and free from distractions. The videorecordings of the speech samples were presented in a large format on one single Qualtrics page without the rating scales. Participants could not pause, stop, replay, or download the videos. Immediately after the video ended, raters were taken to a second page with the rating scales and instructed to rate the video. The stimuli and scales were presented on separate pages to reduce distractions (e.g., rating while listening but not watching), as we wanted the raters to pay close attention to the videos during the entirety of the performance. After the second day, raters completed a short follow-up survey that served to monitor any technical issues with the system.

Data analysis

When preparing the dataset for analysis, the polarity of the scale scores was realigned so that negative judgements (e.g., low comprehensibility, weak grammar, anxious) had an

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

endpoint of 1, while positive judgements (e.g., comprehensible, strong grammar, at ease) aligned with an endpoint of 7. We first calculated polychoric correlations between the scales to determine associations across variables using the *polychoric* function in the *psych* package (version 2.0.8) in R. Polychoric correlations were the basis of analysis because Pearson correlations may attenuate relationships amongst ordinal or ordered categorical variables (Winke et al., 2023), and polychoric correlations provide a more accurate representation of the data (Holgado-Tello et al., 2010). We checked the stability of the correlation matrix against a multilevel Pearson correlation matrix of the same data (reported in Supplement 6), and we found that the single-level correlation matrix was robust for this analysis.

Due to the large number of variables, we then ran exploratory factor analysis (EFA) to determine whether the dataset could be reduced for regression analysis. We ran EFA rather than PCA because component scores are less interpretable than factor scores (Tabachnik & Fidell, 2013), and we hypothesized that the variables would likely show a factor structure due to their semantically related nature (e.g., happy and warm share similar connotations). We first verified that assumptions were met for factor analysis. The relationships between variables were linear, and variance inflation factors were below 4, which satisfied the assumption of a lack of multicollinearity. The Bartlett test of sphericity ($p < .001$) and Kaiser–Meyer–Olkin measure of sampling adequacy (all values > .90) indicated factorability. We used parallel analysis to determine the number of factors rather than eigenvalues greater than 1, as parallel analysis tends to be less biased (Franklin et al., 1995). Parallel analysis indicated the presence of four factors. We then used exploratory factor analysis with the polychoric correlation matrix using maximum likelihood estimation and a promax rotation to produce a factor solution. We used the polychoric correlation matrix because all variables were ordinal (Holgado-Tello et al., 2010). We used an oblique rotation rather than an orthogonal rotation to allow factors to correlate.

Because the dataset included nested data (each participant rated the same 30 samples, thus each participant's scores would exhibit correlations), there was a risk that the factor structure may vary for each rater participant. We were unable to find suitable factor analytic solutions that take into account multilevel data that allow random effects, and pursuing confirmatory factor analysis or structural equation modelling (in which this is possible) was beyond the scope of this study. However, to verify the invariance of the computed factor structure, we bootstrapped the factor analysis with 1,000 samplings of 50 participants from the total pool. The full procedure we followed is detailed in Supplement 6. We found that the factor solution was stable in 97.2% of the bootstrapped calculations, and thus we concluded that the factor structure was stable for this sample despite the multilevel nature of the dataset. Using factor analysis on the full dataset, we extracted factor scores using the Ten Berge method in the *factor.scores* function in R. The Ten Berge method, which is the default method in R, minimizes residuals, produces unbiased estimates of the factor loadings, and is one of the more interpretable methods of producing factor scores (Ten Berge & Kiers, 1991). This method represented the dataset better than orthogonal extraction methods such as the Anderson method, as this method would leave factors uncorrelated, which would have misrepresented the dataset. We then used the three affect-related factor scores in the following regression analysis as independent variables.

To determine how the factor scores related to the language judgements, we built four cumulative logit mixed effects models with the language judgements as separate independent variables. We used cumulative logit mixed effects models rather than multilevel generalized ordered logit models or multilevel ordinal probit regression

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

because we were interested in the overall effects of the predictors rather than modelling the effects at individual thresholds. Cumulative logit models are more parsimonious and can aid in interpretability. The *clmm* function from the *ordinal* package (v.2019.12–10) in R was used for modelling. Random effects of both the rater and sample were entered into the models to account for these sources of variance. We entered all main effects at once, and we used a logit link flexible threshold for all models. We tested the model with main effects against the null model and the same model with random effects removed to ensure that accounting for these sources of variance was meaningful. We applied Bonferroni corrections to the significance threshold to account for the four sets of analyses, with $\alpha = .0125$. All assumptions for these models were met apart from the assumption of proportional odds, which could only be tested using Brant's tests (Brant, 1990) on models with random effects removed using *polyr*. The proportional odds assumption held for some but not all of the main effects in each model. Harrell (2020) stated that in this case, a violation of the proportional odds assumption is not necessarily problematic as long as the focus of the study is to observe average odds ratios for main effects. This was indeed the case in this study, and we thus used the ordinal models instead of less parsimonious multinomial regression models.

The code sheet, written in R, and the dataset are available for analysis in Supplement 4 and Supplement 5 in the Open Science Framework (Burton & Winke, 2025).

Results

The scales showed desirable score distributions, as shown in Table 3. Median scores were slightly more positive than negative for all the categories except grammar and anxiety. Figure 2 shows the distribution of the scale scores across the seven score categories, which indicates that participants tended to avoid the most negative end of the scale. These distributions also showed that the scales were used in different ways, as they have varied patterns. Table 3 also indicates the reliability of the scales. The scales were highly reliable according to Cronbach's alpha calculations (.97–.99), though this reliability is inflated due to a large number of second-level observations (e.g., participants). The intraclass correlation coefficient (ICC), an indicator of interrater consistency, showed a moderate to low amount of consistency, which was anticipated due to the participants' lack of expertise and rater training. This could be variance inherent in

Table 3. Scale means, SDs, and reliability

Scale	Median	Mean	*SD*	Alpha	ICC
Fluency	5	4.58	1.65	.99	.57
Vocabulary	5	4.33	1.67	.99	.54
Grammar	4	4.14	1.59	.98	.37
Comprehensibility	5	4.83	1.64	.99	.48
Engagement	5	5.33	1.31	.98	.40
Anxiety	4	4.12	1.54	.97	.30
Confidence	5	4.45	1.59	.99	.47
Warmth	5	4.76	1.34	.98	.42
Attention	5	5.34	1.23	.98	.35
Expressiveness	5	4.56	1.57	.98	.44
Happiness	5	4.77	1.32	.99	.46
Competence	5	4.87	1.63	.99	.52
Interactiveness	5	4.98	1.39	.98	.39
Attitude	5	5.05	1.22	.98	.40

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

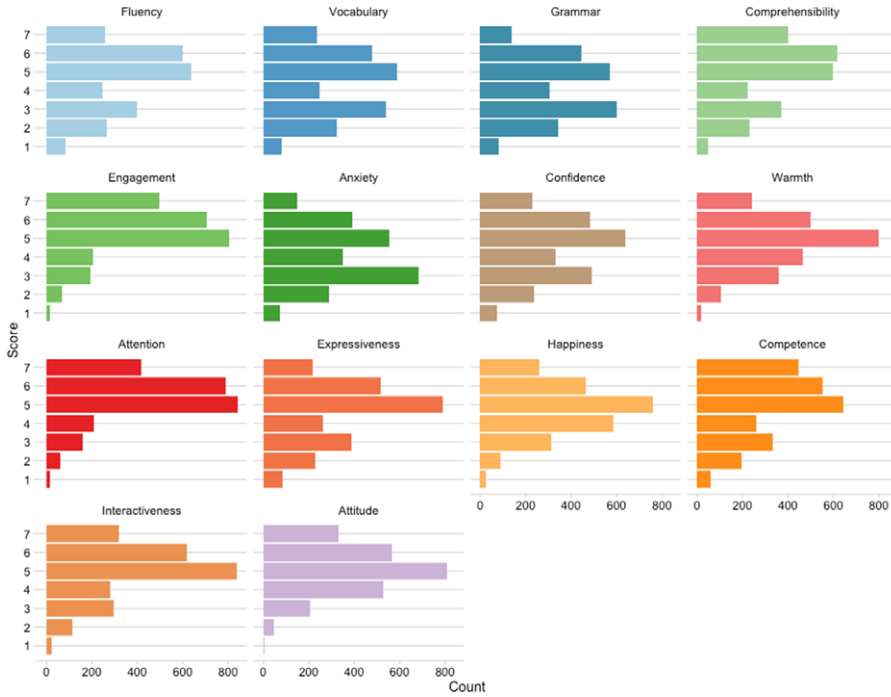

Figure 2. Distribution of scale scores.

affect perception as well, which may be more variable than fixed language features. Notably, ICCs for language-related elements were generally higher, which suggests that raters had a stronger shared intuitive understanding of these characteristics.

The polychoric correlations amongst the scales are presented in Table 4. All correlations were positive, ranging from medium (.40) to strong (≥ .60) (Plonsky & Oswald, 2014). Anxiety and attention correlated the weakest (.40), while fluency and vocabulary correlated the strongest (.85). Particular groups of scales tended to correlate strongly together, such as language elements and competence, features indicating presence (engagement, attention, interactiveness), and positive emotions (attitude, warmth, happiness). However, the ratings across similar features were not identical despite similar scale wording (e.g., happiness, warmth), as they did not exhibit collinearity, with statistical tests showing variance inflation factor (VIF) statistics lower than 4. These relationships suggested that further analyses might be more interpretable using factor scores rather than individual scale categories.

Parallel analysis indicated a 4-factor solution. Figure 3 is a graphical representation of the factor structure in a path diagram (produced using the *psych* package in R, v. 2.4.1). Although these diagrams are more common in confirmatory factor analysis, these diagrams can be useful in EFA to demonstrate relationships among factors graphically (Revelle, 2024). The left boxes represent the 14 scales, while the right circles are the factors that predict scale values. The constrained factor loadings (fixed to the strongest loading factor) are represented on the arrows from the latent factor to the scale. These represent the relative strength of the loadings rather than absolute loadings. The correlations between the factors are shown on the far–right–hand side. The factor structure was anticipated from the correlation data, though we did not anticipate that

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

Table 4. Scale 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

competence would be more bound to language features than affect features. In this model, we renamed the factors using terms that most aligned with their apparent meaning: language (fluency, vocabulary, grammar, comprehensibility, and competence), (self-) assuredness (confidence and anxiety), involvement (engagement, attention, and interactiveness), and positivity (happiness, warmth, attitude, and expressiveness). We then extracted factor scores of the affect factors to represent assuredness, involvement, and positivity for regression analysis as predictors of the original observed variables: fluency, grammar, vocabulary, and comprehensibility.

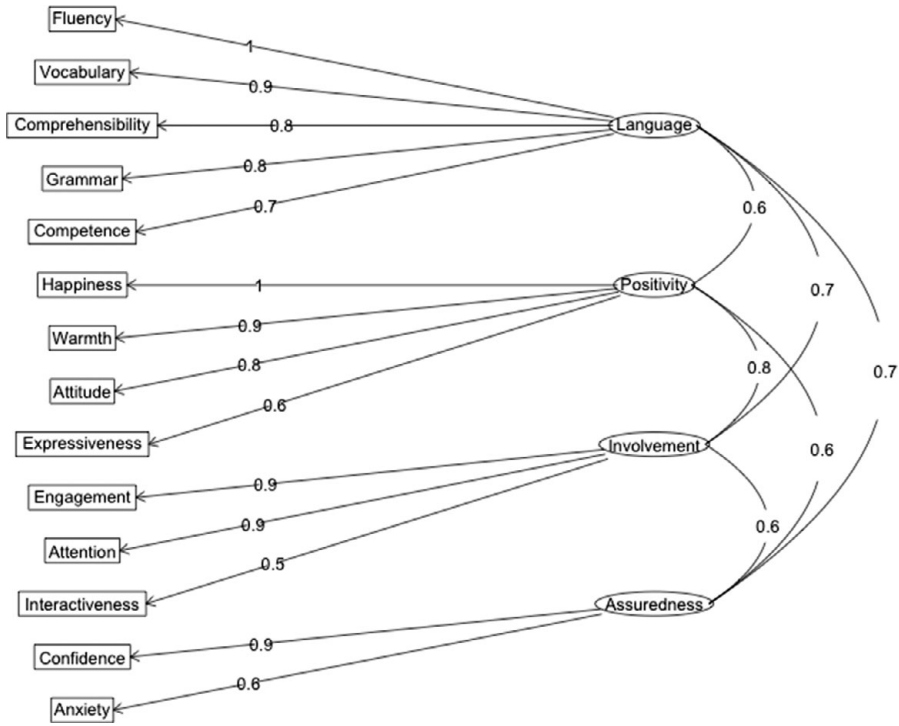

Figure 3. Path diagram of factor solution.

Table 5. Polychoric correlations between factor scores and language scores

Scale	Assuredness	Involvement	Positivity
Fluency	.72 [.69, .74]	.67 [.65, .70]	.59 [.55, .61]
Vocabulary	.68 [.65, .71]	.61 [.59, .64]	.54 [.51, .57]
Grammar	.59 [.56, .63]	.52 [.49, .55]	.48 [.44, .51]
Comprehensibility	.62 [.59, .65]	.63 [.61, .66]	.60 [.57, .63]

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

Relationships between affect and spoken language ability

The polychoric correlations between the factor scores and the four original language scores were all positive and moderate to strong, ranging from .48 to .72, as shown in Table 5. Each language feature was more strongly associated with assuredness and involvement than positivity. These associations were stronger for fluency than the other language scores. Regarding positivity, it showed the strongest correlation with comprehensibility (.60). The weakest associations were with grammar.

We built four sets of mixed effects ordinal regression models to determine which of the factor score reductions of the affect variables had the greatest impact on each score category after taking into account the variance attributable to each participant and each speech sample. In these models, the model with main effects fit better than the null model with main effects and no predictors (null model 1) or the model with main effects with random effects removed (null model 2), as shown in Table 6. The models of fluency (Table 7), vocabulary (Table 8), and grammar (Table 9) showed similar patterns in their main effects. In each of these models, only assuredness and involvement were significant predictors of each language score, with the strongest associations between assuredness and fluency, β = .81, odds ratio = 2.25, and involvement and fluency, β = .79, odds ratio = 2.22. These indicate that the likelihood of a score category increase on the 7-point scale in fluency, vocabulary, and grammar was roughly two times greater with 1-point increases in perceptions of assuredness and involvement. These models explained roughly a fifth to a quarter of the variance in each model using Nagelkerke's Pseudo R^2, fluency = .27, vocabulary .23, grammar = .18.

Table 6. Tests of model fit

Category	Model	Likelihood Ratio	df	*p*-value	AIC	BIC
Fluency	Null 1				7016.70	7063.23
	Null 2	105.83	1	< .001	6912.80	6965.22
	Main	634.25	2	< .001	6282.60	6346.62
Vocabulary	Null 1				7325.30	7271.87
	Null 2	–19.432	1	> .99	7346.70	7399.13
	Main	619.96	2	< .001	6730.80	6794.81
Grammar	Null 1				7765.20	7811.80
	Null 2	23.29	1	< .001	7743.90	7796.33
	Main	441.67	2	< .001	7306.30	7370.30
Comprehensibility	Null 1				7246.90	7293.43
	Null 2	–48.26	1	> .99	7297.10	7349.53
	Main	633.22	2	< .001	6667.90	6731.94

Table 7. Fluency model

Coefficients	β	95% CI	*SE*	*z*	*p*	OR	95% CI
Assuredness	.81	[.72, .90]	.05	16.98	< .001	2.25	[2.05, 2.47]
Involvement	.79	[.66, .93]	.07	11.68	< .001	2.22	[1.94, 2.53]
Positivity	–.03	[–.16, .11]	.07	–.38	.70	.97	[.85, 1.12]
	Random effects						
	Groups	Variance	SD				
	Raters	.69	.83				
	Samples	1.65	1.29				

Note: $p < .0125$.

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

Table 8. Vocabulary model

Coefficients	*β*	95% CI	*SE*	*z*	*p*	OR	95% CI
Assuredness	.70	[.59, .77]	.05	14.75	< .001	1.97	[1.80, 2.16]
Involvement	.68	[.55, .82]	.07	10.19	< .001	1.98	[1.74, 2.26]
Positivity	.002	[–.13, .14]	.07	.02	.98	1.00	[.88, 1.14]
	Random effects						
	Groups	Variance	SD				
	Raters	.53	.73				
	Samples	1.44	1.20				

Note: $p < .0125$.

Table 9. Grammar model

Coefficients	*β*	95% CI	*SE*	*z*	*p*	OR	95% CI
Assuredness	.59	[.50, .67]	.04	13.15	< .001	1.80	[1.65, 1.96]
Involvement	.54	[.42, .67]	.06	8.35	< .001	1.72	[1.51, 1.95]
Positivity	.04	[–.09, .17]	.07	.60	.55	1.04	[.91, 1.19]
	Random effects						
	Groups	Variance	SD				
	Raters	.67	.82				
	Samples	.61	.78				

Note: $p < .0125$.

Table 10. Comprehensibility model

Coefficients	*β*	95% CI	*SE*	*Z*	*p*	OR	95% CI
Assuredness	.51	[.41, .60]	.05	11.27	< .001	1.66	[1.52, 1.82]
Involvement	.58	[.45, .71]	.07	8.74	< .001	1.78	[1.57, 2.03]
Positivity	.36	[.23, .50]	.07	5.31	< .001	1.44	[1.26, 1.65]
	Random effects						
	Groups	Variance	SD				
	Raters	.85	.92				
	Samples	1.11	1.06				

Note: $p < .0125$.

The model for comprehensibility, shown in Table 10, differed slightly. In this model, all three main effects were significant, with somewhat smaller standardized coefficients and odds ratios. This model deviated from the previous models in that the relationship between involvement (rather than assuredness) and the outcome was the strongest, though only slightly (odds ratio for involvement = 1.78, assuredness = 1.66). As opposed to the previous three models, positivity was a significant, positive predictor of comprehensibility, $\beta = .36$, odds ratio 1.44, showing that the likelihood that speakers were classified as one point easier to understand on the 7-point scale was about 1.44 times higher when there were 1-point increases in perceived displays of positive affect. The model for comprehensibility explained about one fifth of the variance in the comprehensibility scores, Nagelkerke's Pseudo $R^2 = .22$.

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

Discussion

The goal of this study was to investigate whether and to what degree perceived affect relates to spoken language ability judgements. This study showed that listeners' perceptions of L2 speakers' extra-linguistic, affect displays (e.g., emotions and social orientations) interweave with their judgements of L2 speech to a sizable degree. Correlations showed that all ten measures of affect (e.g., confidence, engagement, warmth) correlated positively with all four linguistic measures with varying levels of strength. This fact suggests that spoken language ability as conceived by linguistic laypeople is a complex construct that may consist of multiple nonverbal, contextual elements drawn from the visual world. If this is indeed true, it can partially explain differences in how language is perceived across modalities, where having access to audiovisual content over audio alone tends to result in stronger perceptions of comprehensibility and language proficiency (Carey & Szocs, 2024; Nakatsuhara et al., 2021). Nonetheless, upon closer inspection, not all measures of affect corresponded to language ratings equally, as there was nuance in which measures of affect were most likely to relate to certain domains of language.

Because the affect measurements tended to cluster together in their correlations, we extracted three factors to investigate broader relationships between affect and L2 speech. We named these factors assuredness (confidence and anxiety), involvement (engagement, attention, and interactiveness), and positivity (happiness, warmth, expressiveness, and attitude). Competence, contrary to our expectations, was perceived as a language judgement, which the raters may have used as a proxy measure of listening comprehension. We found that assuredness had the strongest relationship with fluency, vocabulary, and grammar scores; in these cases, when individuals are seen as being more confident and at ease (assured), they are more likely to be perceived as stronger in each of these three linguistic areas. This finding is largely in line with Clément (1986) and Noels and Clément (1996), who argued that confidence (especially self-confidence as reported by the speaker) was one of the strongest predictors of language proficiency. Likewise, it is also in line with the vast literature on anxiety, which has found that low anxiety may relate to positive proficiency or achievement outcomes (e.g., Botes et al., 2020; MacIntyre et al., 1997; Teimouri et al., 2019). Confidence is an affective stance that raters frequently observe and factor into positive evaluations of test-takers in the language testing literature as well (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 and affective cues to extrapolate information about the test takers' underlying cognitive fluency (ability to process language quickly and efficiently) and lexicogrammatical competence. Seeing an individual as anxious or less confident may have led raters to perceive that person as less proficient, resulting in lower scores being awarded. Likewise, seeing a confident performance could inform raters that the speaker believed in their own abilities, thus leading to higher gains. Perceptions of confidence may be bidirectional, however; that is to say, people may find more confident and less anxious speakers to be more proficient overall, but more proficient individuals are likely to be perceived as more confident and at ease simply based on their stronger language skills (Edwards & Roger, 2015). What stands out in this study is that these relationships existed external to the speaker in the eyes of listeners rather than through self-reports of affect and language ability as in previous research.

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

Involvement, made up of cognitive, social, and perhaps behavioural engagement in the speaking scenario, also stood out as a strong predictor of the same three linguistic outcomes, though slightly less so than assuredness did. Raters may have found that speakers who were more attentive and interactive with the examiner were able to display a greater range of evidence of their spoken language ability. This may have also been the case during question breakdowns, where speakers may have shown more engagement by asking follow-up questions to repair the breakdown sequences. That these displays of involvement were associated with stronger perceived language in fluency, grammar, and vocabulary is supported by findings that engagement can lead to greater task success (Storch, 2008) as well as positive impressions of communicative and interactional competence (Ducasse & Brown, 2009; May, 2011; Sato & McNamara, 2019). Other studies have also found links between willingness to communicate, which may be closely related to involvement, and L2 (communicative) competence (Elahi Shirvan et al., 2019; Jin & Lee, 2022).

Interestingly, emotional components aligned with positivity were not components of the factor structure of involvement/engagement, despite past theorizations (Philp & Duchesne, 2016), though these two factors did correlate strongly (.80). Past research has found that foreign language enjoyment, which may manifest as positive affect in discrete scenarios, may correspond with achievement in an L2 (Botes et al., 2020; Dewaele & Li, 2022; Li et al., 2020). Similarly, one recent study found that smiling and laughing (behaviours closely related to positive affect) were associated with judgements of greater fluency with a correlation of .42 (Kim et al., 2024). Although the current study found similar strength in correlations with positive affect (fluency = .59, vocabulary = .54, grammar = .48), positivity did not emerge as a predictor of fluency, grammar, or vocabulary in any of the models in this study. It may be the case that one's own perceived enjoyment (as a longer-term trait) may be a motivating factor in overall language acquisition, but temporary displays of happiness or warmth in an interactive context may exert less of an effect on perceived ability, especially when other affective phenomena such as assuredness and involvement are considered.

The judgements of comprehensibility in this study showed somewhat different patterns from fluency, vocabulary, and grammar. Although both assuredness and involvement predicted comprehensibility as well, involvement emerged as a slightly stronger predictor of the two. Positivity, in contrast to the previous models, also predicted comprehensibility. All aspects of spoken performance as measured in this study benefitted from perceived assuredness and involvement, and this finding with comprehensibility is supported by past research. In Nagle et al. (2022), for example, both low anxiety and collaborativeness, a measure of social engagement, predicted comprehensibility outcomes in pairs of dyads. Novel in the current study is that positive affect was also a component of the variance in these scores. There is some support for this finding from recent literature on nonverbal behaviour, showing that looking away (possibly indicating content-related thinking), smiling, and backchanneling with head nods may correspond with comprehensibility judgements (Tsunemoto et al., 2022; Trofimovich et al., 2021), but the benefit derived from behaviours with a positive valence may only apply to learners with lower proficiency (Burton, 2024). Although features of speech prosody (such as intonation) may exhibit variable relationships with comprehensibility across ability levels (Huensch & Nagle, 2021; Kang et al., 2010; Munro & Derwing, 1999; Sereno et al., 2016; Trofimovich & Isaacs, 2012), intonational contours have been found to relate strongly to certain vocal emotions, such as positivity (Larrouy-Maestri et al., 2024; Rodero, 2011). It is possible that paralinguistic prosodic features indirectly enhanced comprehensibility through perceived positive affect along

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

with nonverbal facial behaviours. Given that the broader construct of engagement is also made up of components that relate to positive affect (Philps & Duchesne, 2016), this may indicate that the relative ease of understanding of speakers is partially comprised of how pleasant, interactive, collaborative, and at ease speakers appear to listeners in addition to the myriad linguistic factors that have been documented (Crowther et al., 2016; Isaacs & Trofimovich, 2012). Overall, these various findings suggest that a wide range of linguistic and extra-linguistic factors interact when listeners are decoding second-language speech.

We have speculated on the mechanisms that may drive relationships between assuredness and involvement in these measures, but how positive affective variables impact ease of understanding is unclear. One possible explanation could reside within the literature on affective or emotional contagions (Elfenbein, 2014; Smirnov et al., 2019). Affective contagions are emotions that are contracted by an interactant when they unconsciously and automatically converge on an undirected emotion (an emotional display without a clear source). In other words, the feeling spreads from speaker to listener. In this context, when a listener sees a speaker exhibiting positive affect (and possibly other affective orientations such as engagement), research suggests that some of that affect transfers to the listener, making them feel more invested in paying attention and listening to the speaker. In other words, the speaker's stances may inspire a *willingness to listen* in their interactant. Listeners who become more willing to listen to a particular speech sample may likewise find it easier to understand because of the increased effort in decoding speech. Being perceived as easier to understand could then have corresponding benefits for how other aspects of language, such as vocabulary and grammar, are perceived by listeners.

Implications

This study has theoretical implications for how spoken language ability is conceptualized. Models of language ability or communicative competence (e.g., Bachman & Palmer, 1996; Canale & Swain, 1980) describe second-language communication almost entirely in terms of how linguistic components of speech and writing convey meaning in sociocultural contexts. Affect and behaviour are generally not considered to be major components in these models apart from minor compensatory roles in strategic competence (Canale & Swain, 1980). Models of interactional competence (e.g., Galaczi & Taylor, 2018) are more inclusive of behaviour and affect, especially their role in turn-taking and conversation management, but these models do not explain how affect may closely interact within a broader communicative system. Hymes's (1972) original concept of communicative competence, however, allowed for affective orientations to interact with language ability through the *ability for use*, or the functional ability of an individual to deploy their linguistic competence. The findings in this study that assuredness and involvement closely overlap with judgements across all rated categories of speech appear to support a model which includes ability for use, suggesting that these models may need closer inspection in the future. Nonetheless, the exact mechanisms of how affect relates to the various competencies in such a model need more attention before further conclusions are drawn.

These results also have implications for research and practice. For one, although constructs such as comprehensibility have received ever more attention in recent literature, the complex relationship between affect and ease of understanding complicates its measurement if visual content is included. Researchers may need to consider or control for elements of engagement (e.g., backchanneling, mutual gaze) and positive affect (e.g., smiling, laughing), as these elements appear to be able to enhance second

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

language speech comprehension. In terms of pedagogy, there may be a question of whether behavioural elements of positive affect should be taught or encouraged in the language classroom with the belief that this could result in beneficial outcomes for the speaker. There is the possibility that this may be true in the cultural context of this study (the United States), as these types of social and psychological behaviours are common in service encounters such as when dining out, or within sensitive or high-stakes communicative contexts (health care settings, negotiations with law enforcement, diplomatic negotiations or even when one is asking for homework to not be counted as late). However, it may also be the case that the detection of *stilted* affect (unnecessarily formal affect) could backfire. Studies considering natural or forced affect could be revealing as to whether there is any pedagogical value in encouraging certain behaviours or affect displays.

For language assessment practice, this study raises key considerations about the fairness of including or ignoring affect in proficiency evaluations where test takers are visible to examiners. One would have serious doubts about the fairness of a test if, for example, a candidate who smiled less received a lower test score. Language tests are stressful and marked by unequal distributions of power between the examiner and test taker, which could very well result in more serious-appearing behaviour. Positive affect in this study, however, did not predict changes in fluency, grammar, or vocabulary scores when other measures of affect were considered, which appears to support fairness regarding this particular measurement. Even though this study did not find that positive affect related to these linguistic measures, certain populations may exhibit different patterns of behaviours or affect depending on group or individual differences. For example, research has shown that culture mediates how nonverbal behaviour may be encoded by speakers and decoded by listeners (Matsumoto & Hwang, 2016), and even emotional affect is a culturally constructed phenomenon (Mesquita, 2022). If there are behavioural norms in one culture (e.g., avoiding eye contact with superiors) that are seen as less appropriate in another (e.g., eye contact is seen as polite), there may be effects that arise when individuals from these cultural backgrounds perceive each other's speech. This could result in unfair or biased test scores if certain groups receive higher scores simply because their cultural norms match those expected by the raters. Likewise, other individuals exhibit neurological differences that may conflict with what is seen as "standard" in the broader population. Test takers with autism, for example, may avert their gaze more and produce repetitive motions (American Psychiatric Association, 2013) while producing speech that may be identical or equivalent to neurotypical test takers. In test-taker populations that include neurodiverse individuals or groups from different cultures, investigations of behavioural bias on test scores are critical. If bias is found to exist, steps should be taken to ameliorate this. For neurodiverse test takers, this may mean providing accommodations or modifications to how the test is rated that ensure fairness.

The fairness of considering assuredness and involvement in test scores, as this study showed, is more complex. If indeed assuredness is a bidirectional result of language proficiency, and if assuredness is partially a cognitive mechanism (e.g., Stankov et al., 2012), being perceived as more confident or less anxious may in fact reveal something about an individual's L2 ability in *some* cases. Testing situations are anxiety-producing, though, and the mere fact of feeling anxious or having lower confidence may have more to do with that person's personality than underlying ability. Showing nonverbal evidence of engagement, attention, and interactiveness, however, are all important skills in effective L2 communication, especially within subdomains such as interactional competence (Galaczi & Taylor, 2018; Plough et al., 2018) or goal-directed communicative

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

effectiveness (Morreale et al., 2013). Displaying this type of affect alone is not enough to overcome very low proficiency, but it may help facilitate intercultural encounters. Nevertheless, while these affective responses may provide implicit information for raters about underlying spoken language ability, scale development and rater training programs should be investigated to ensure fairness in these measurements as well, especially when working with diverse test-taking populations (Randez & Cornell, 2023).

Limitations

This study naturally comes with several limitations. Notably, using untrained, naïve raters results in a large amount of variance in how outcomes are perceived, and part of this variance may be due to differing internal definitions of each of these language categories. For example, even though a narrow definition of fluency was provided to the participants to differentiate this from the broader definition generally used in society (Lennon, 1990), follow-up qualitative data showed that some participants oriented to fluency as a general measure of language proficiency (see Burton, 2023). The same held for the term comprehensibility, which was occasionally discussed in relation to comprehension, though this was rare. These varying internal representations of facets of language could have skewed or attenuated some of the relationships found with affect variables. However, modelling raters as a random effect in the mixed effects model accounts for these varying patterns and thus strengthens the inferences from the regression models reported in this study. Future work should consider whether these effects exist even in pools of trained language professionals using more descriptive rating scales.

Similarly, even though pilot testing showed sufficient separability in the constructs being measured, there is always a risk of raters marking diverse rating criteria *too* similarly, known as the *halo effect*. The halo effect can muddy boundaries between otherwise distinct constructs. This type of effect may be responsible for at least part of the variance in the correlations reported in this study, though we note that the distinctive clustering patterns in the factor models indicate that the participants viewed these criteria as sufficiently different. Future work could separate counterbalanced rating sessions by affect and language ability to determine whether these relationships still stand when these varying constructs are observed on different occasions. Likewise, like the previous point, training raters on more descriptive language ability criteria would also serve to reduce potential halo effects and result in stronger inferences about the rated constructs.

Finally, a limitation and simultaneously a strength of the study was the controlled nature of the backgrounds of both the L1 raters (US-born English speakers) and the L2 test takers (Chinese-born speakers of one or more dialects of the Chinese language). Controlling for the participants' backgrounds helped to isolate the effects of nonverbal behaviour and prevent some covariance due to culture (although regional variation in affect displays and their interpretation may have been present in the dataset). These controlled backgrounds did not allow us to measure how the impact of assuredness, involvement, and positivity might have remained invariant depending on different L1 and L2 groups. There is some indication from the literature that it may not (Uchida et al., 2009; although cf. Tsunemoto et al., 2022). We do not know if, for example, the L1 American listeners would demonstrate the same relationships between affect and the L2 speech of European, South American, or African test takers. Similarly, we do not know if different L1 groups from, for example, the United Kingdom, India, Australia, or any of the growing spheres of English as a lingua franca would demonstrate the same relationships when watching and listening to the same L2 groups. This is a complex area

https://doi.org/10.1017/S0272263125000063 Published online by Cambridge University Press

of study, so isolating and identifying these effects in various L1–L2 groupings is a promising area of future research.

Conclusion

Second language communication is a complex, multifaceted construct. With increased research tapping into the relationship between L2 linguistic outcomes and various features often perceived in the visual world (Carey & Szocs, 2024; Jenkins & Parra, 2003; Nakatsuhara et al., 2021; Trofimovich et al., 2021; Tsunemoto et al., 2022), a picture is beginning to emerge of language ability as inherently multimodal in the range of features that influence its perception. This has a wide range of implications, especially in the measurement of language ability, as the constructs that underlie these measurements largely do not account for how learners leverage affect to convey meaning. For SLA research that aims to focus purely on linguistic outcomes, affective phenomena could result in noisy statistics. On the other hand, for those interested in more holistic interpretations of how well learners can communicate, not accounting for affect (by, for example, only considering audio recordings or transcriptions) could disadvantage learners as their entire repertoire of communication would not be considered. More research is needed to continue exploring the impact of these variables and the mechanisms that influence language perception. Ultimately, this is a question of fairness, as a greater understanding can help inform more accurate and precise decisions made about learners that determine their access to opportunities in the real world.

Data availability statement. The experiment in this article earned Open Data badge for transparent practices. The data are available at https://url.avanan.click/v2/r02/.

References

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

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Burton, J. D. (2023). *The role of nonverbal behavior and affect on ratings of second language proficiency* [Doctoral dissertation, Michigan State University]. ProQuest Dissertations and Theses Global.

Burton, J. D. (2024). Evaluating the impact of nonverbal behavior on language ability ratings. *Language Testing*, *41*(4), 729–758. https://doi.org/10.1177/02655322241255709

Burton, J. D., & Winke, P. (2025). *Affect as a component of second language speech perception* (OSF 2FTVJ, Version V1) [Data set]. Open Science Framework. https://doi.org/10.17605/OSF.IO/2FTVJ

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language test*. Oxford University Press.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*(7), 462–479. https://doi.org/10.1037/0003-066X.54.7.462

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

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

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

Carey, M. D., & Szocs, S. (2024). Revisiting raters' accent familiarity in speaking tests: Evidence that presentation mode interacts with accent familiarity to variably affect comprehensibility ratings. *Language Testing*, *41*(2), 290–315. https://doi.org/10.1177/02655322231200808

Chong, J. J. Q., & Aryadoust, V. (2023). 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

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

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

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility: Nonnative listeners' perspectives. *Journal of Second Language Pronunciation*, *2*(2), 160–182. https://doi.org/10.1075/jslp.2.2.02cro

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

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

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

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

Elahi Shirvan, M., Khajavy, G. H., MacIntyre, P. D., & Taherian, T. (2019). A meta-analysis of L2 willingness to communicate and its three high-evidence correlates. *Journal of Psycholinguistic Research*, *48*, 1241–1267. https://doi.org/10.1007/s10936-019-09656-9

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

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., & Fralish, J. S. (1995). Parallel analysis: a method for determining significant principal components. *Journal of Vegetation Science*, *6*(1), 99–106. https://doi.org/10.2307/3236261

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.

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

Gluszek, A., & Dovidio, J. F. (2010). Speaking with a nonnative accent: Perceptions of bias, communication difficulties, and belonging in the United States. *Journal of Language and Social Psychology*, *29*(2), 224–234. https://doi.org/10.1177/0261927X09359590

Harrell, F. (2020, September 20). *Violation of proportional odds is not fatal.* Statistical Thinking. https://www.fharrell.com/post/po

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., & Vila–Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *44*, 153–166. https://doi.org/10.1007/s11135-008-9190-y

Hox, J., Moerbeck, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge. https://doi.org/10.4324/9781315650982

Huensch, A., & Nagle, C. (2021). The effect of speaker proficiency on intelligibility, comprehensibility, and accentedness in L2 Spanish: A conceptual replication and extension of Munro and Derwing (1995a). *Language Learning*, *71*(3), 626–668. https://doi.org/10.1111/lang.12451

Hymes, D. (1972). On communicative competence. In A. Duranti (Ed.), *Linguistic anthropology: A reader* (pp. 53–73). Blackwell.

IELTS. (n.d.). *IELTS.* https://ielts.org

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*(3), 475–505. https://doi.org/10.1017/S0272263112000150

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

Jin, S., & Lee, H. (2022). Willingness to communicate and its high-evidence factors: A meta-analytic structural equation modeling approach. *Journal of Language and Social Psychology*, *41*(6), 716–745. https://doi.org/10.1177/0261927X221092098

Kang, O., & Yaw, K. (2024). Social judgement of L2 accented speech stereotyping and its influential factors. *Journal of Multilingual and Multicultural Development*, *45*(4), 544–551. https://doi.org/10.1080/01434632.2021.1931247

Kang, O., Rubin, D. O. N., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, *94*(4), 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

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

Kim, Y. L., Liu, C., Trofimovich, P., & McDonough, K. (2024). Is nonverbal behavior during conversation related to perceived fluency? *TESOL Journal*. Article e795. Advance online publication. https://doi.org/10.1002/tesj.795

Larrouy-Maestri, P., Poeppel, D., & Pell, M. D. (2024). The sound of emotional prosody: Nearly 3 decades of research and future directions. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/17456916231217722

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

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

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., Babin, P. A., & Clément, R. (1999). Willingness to communicate: Antecedents & consequences. *Communication Quarterly*, *47*(2), 215–229. https://doi.org/10.1080/01463379909370135

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. D., 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

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

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

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

Mesquita, B. (2022). *Between us: How cultures create emotions*. W. W. Norton & Company.

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

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, *49*(s1), 285–310. https://doi.org/10.1111/0023-8333.49.s1.8

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., Inoue, C., & Taylor, L. (2021). 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

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.

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

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

Pavlenko, A. (2014). *The bilingual mind: And what it tells us about language and thought*. Cambridge University Press. https://doi.org/10.1017/CBO9781139021456

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

Ploder, A., & Eder, A. (2015). Semantic differential. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences*, 2nd Ed. (pp 563–571). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.03231-1

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., Banerjee, J., & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing*, *35*(3), 427–455. https://doi.org/10.1177/0265532218772325

Randez, R. A., & Cornell, C. (2023). Advancing equity in language assessment for learners with disabilities. *Language Testing*, *40*(4), 984–999. https://doi.org/10.1177/02655322231169442

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research.* R package version 2.4.1. https://CRAN.R-project.org/package=psych

Rodero, E. (2011). Intonation and emotion: influence of pitch levels and contour type on creating emotions. *Journal of Voice*, *25*(1), e25–e34. https://doi.org/10.1016/j.jvoice.2010.02.002

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

Sereno, J., Lammers, L., & Jongman, A. (2016). The relative contribution of segments and intonation to the perception of foreign-accented speech. *Applied Psycholinguistics*, *37*(2), 303–322. https://doi.org/10.1017/S0142716414000575

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

Smirnov, D., Saarimäki, H., Glerean, E., Hari, R., Sams, M., & Nummenmaa, L. (2019). Emotions amplify speaker-listener neural alignment. *Human Brain Mapping*, *40*(16), 4777–4788. https://doi.org/10.1002/hbm.24736

Snider, J. G., & Osgood, C. E. (Eds.). (1969). *Semantic differential technique: A sourcebook*. Aldine.

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

Storch, N. (2008). Metatalk in a pair work activity: Level of engagement and implications for language development. *Language Awareness*, *17*, 95–114. https://doi.org/10.1080/09658410802146644

Tabachnik, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

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

Ten Berge, J. M., & Kiers, H. A. (1991). A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika*, *56*, 309–315. https://doi.org/10.1007/BF02294464

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., 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

Winke, P., Zhang, X., & Pierce, S. (2023). A closer look at a marginalized test method: Self-assessment as a measure of speaking proficiency. *Studies in Second Language Acquisition*, *45*(2), 416–441. https://doi.org/10.1017/S0272263122000079

Cite this article: Burton, J. D., & Winke, P. (2025). Affect as a component of second language speech perception. *Studies in Second Language Acquisition,* 1–26. https://doi.org/10.1017/S0272263125000063