L1 listeners' evaluations of LX speech: the role of listener expectations and personality traits

Hui Sun, Petar Čolović, Petar Milin & Dagmar Divjak

Published online: 29 Oct 2024.

Submit your article to this journal ↗

View related articles ↗

View Crossmark data ↗

Routledge
Taylor & Francis Group

L1 listeners' evaluations of LX speech: the role of listener expectations and personality traits

Hui Sun[a,b], Petar Čolović[c], Petar Milin[a] and Dagmar Divjak[a,d]

[a]Department of Modern Languages, University of Birmingham, Birmingham, UK; [b]School of English, Communication and Philosophy, Cardiff University, Cardiff, UK; [c]Department of Psychology, University of Novi Sad, Novi Sad, Serbia; [d]Department of Linguistics & Communication, University of Birmingham, Birmingham, UK

ABSTRACT

LX speech tends to receive less favourable judgments, which can be attributed to its non-standard linguistic properties (e.g. foreign accents, grammatical errors). However, the evaluations of LX speech could also be affected by listeners' expectations and their individual differences. This study extends research on the role of listeners' expectations (cued by speech accent) and personality traits (Big Five), by examining the degree to which and how accent, grammaticality and listener personality interact to determine L1 listeners' evaluations of LX speech. Sixty L1 English speakers (30 female) rated the acceptability of 40 English speech stimuli in a British or Polish accent; half of the stimuli were filled with English article errors. The listeners' personality traits were self-reported via a short questionnaire of the Big Five factors. The significant interaction of accent, grammaticality and listener personality suggests that the occurrence of grammatical errors is less penalised in foreign-accented speech (vs. native-accented speech), unless the L1 listeners are highly conscientious or highly extraverted. The theoretical (expectation mismatch effect, Interpersonal Circumplex model), methodological (considering non-linear models and the moderation of contextual factors for personality effects) and potential practical implications (awareness training of personality effects) of the results are discussed.

Introduction

With English being used globally, communication in English in a natural environment often involves both L1 and LX[1] users. This relationship is, however, one of power: deviations from the (local) norm are typically ill-received, such as foreign accents and grammatical errors in LX speech (Baese-Berk, McLaughlin, and McGowan 2020; Eisenstein 1983). The literature on language attitudes has widely documented a negative perception of foreign accents (Dragojevic et al. 2021). For instance, L1 listeners tend to rate foreign-accented speakers lower in status (socioeconomic position) and solidarity (ingroup favouritism) (Fuertes et al. 2012; Nejjari et al. 2012), which could have a negative impact on the evaluations of them in high-stakes contexts, including education, employment, immigration and court (Munro 2003; Romero-Rivas, Morgan, and Collier 2022; Spence et al. 2022). Though less researched, grammatical errors have also been found to lower ratings of 'social'

and 'academic' competence of LX users (Queen and Boland 2015), and tend to be even less acceptable when occurring with a heavy accent (Ruivivar and Collins 2019).

However, the evaluation of LX speech is a complex process, affected not only by the speech itself, but also by the listener' expectations about the speaker's language performance (Lindemann and Subtirelu 2013). Drawing on social information (e.g. cues for social stereotypes), listeners tend to form expectations actively, which could help them with input processing. When there is a mismatch between expected and observed performances (e.g. a Caucasian speaker rather than an Asian speaker producing foreign-accented English speech), listeners tend to have more difficulty in speech processing and give more negative evaluations, possibly due to cognitive incongruence (Gnevsheva 2018; McGowan 2015). So far, research on the role of listener expectations in LX evaluations, through verbal/matched guise tests (see Garrett 2010 for guise techniques), has mainly focused on foreign accent without considering grammatical errors, although these two features co-occur frequently in LX speech. To fill the gap, this study examined the interaction effect of accent (cue for expectations) and grammaticality on L1 listeners' evaluations of LX speech.

Furthermore, the variance in LX speech evaluations has been related to a range of listener individual differences, including age, gender, education level, linguistic metaknowledge, familiarity with foreign-accented speech, and personality (e.g. Dewaele and McCloskey 2015; Kang 2012; Saeli 2018; Tsurutani and Selvanathan 2013). Similarly, these studies examined foreign-accented speech without paying attention to grammaticality. Given that the role of personality as a complex construct requires further exploration, we investigated the degree to which L1 listeners' personality traits interact with accent and grammaticality to determine their evaluations of LX speech.

Listener expectations and language evaluations

Visual cues have been widely used to form the listener's expectations about the speaker's social membership and language ability – a Caucasian face is associated with L1 speakers of English and an Asian face with LX speakers of English (Gnevsheva 2018; Kang and Rubin 2009; McGowan 2015; Rubin 1992). Rubin and colleagues found that when the assumed speaker was Asian rather than Caucasian, listeners perceived native-accented speech to be foreign-accented and less comprehensible. They accounted for the findings by 'reverse linguistic stereotyping', suggesting a top-down perceptual approach where the listener's speech perception is influenced by the social stereotypes (or bias) of the speaker (Kang and Rubin 2009).

McGowan (2015), however, interpreted the findings from the perspective of cognitive congruence drawing on usage-based exemplar models (K. Johnson 1997, 2006). That is, the alignment between listener expectations and speech stimuli facilitates the activation of the relevant exemplars and thus enhances the speech perception, whereas a mismatch inhibits the access to exemplars and degrades the perception. This 'expectation mismatch effect' is supported by McGowan's results that when the speech stimuli fulfilled listeners' expectations (i.e. Chinese-accented speech being presented with a Chinese face rather than a Caucasian face), the intelligibility of the speech stimuli was enhanced. Gnevsheva's (2018) inclusion of an audio-only baseline condition provided further evidence. Compared with the accentedness rating of foreign-accented speech in audio stimuli, the appearance of a Caucasian face in the video stimuli made the speech sound more accented (expectation mismatch), while the appearance of an Asian face did not change the ratings.

One can hypothesise that the detrimental effect of expectation mismatch may be extended to the evaluations of LX speech with grammatical errors. A recent study by Kirjavainen et al. (2024) examined the effect of L1 readers' expectations on their perception of grammatical errors in written texts in this light. They considered eight grammatical errors of the kind produced by LX speakers in English speaking classrooms yet not serious enough to impede meaning, including errors concerning plurality, tense, articles, copula, etc. Readers' expectations were evoked by guise names, with English, Swedish and Chinese names associated with L1, high LX and low LX English ability level respectively. When those grammatical errors occurred, the language ability ratings of English

and Swedish writers significantly decreased, probably because readers expected them to produce relatively error-free texts. In contrast, Chinese writers' language ability ratings were not affected by the occurrence of grammatical errors, as errors were likely expected.

Given that listeners can recognise whether the accent is native or not in as little as 30 ms (Flege 1984) and that LX grammatical errors appear frequently in foreign-accented (but not native-accented) speech, it is reasonable to assume that speech accent prompts the expectation of grammatical errors. Indeed, there is evidence in neurolinguistic research that grammatical errors in native-accented speech tend to trigger more pronounced neural responses than those in foreign-accented speech (Grey and van Hell 2017; Hanulíková et al. 2012). Although Ruivivar and Collins (2018, 2019) examined the acceptability of grammatical errors in different accent conditions, the role of listener expectations is difficult to discern due to the missing baseline (error-free) condition. Thus, this study tests whether L1 listeners would lower the ratings of LX speech to a larger extent when grammatical errors occur in native-accented speech (expectation mismatch), compared with when they hear these same errors in foreign-accented speech.

Personality traits and language evaluations

While language evaluations at the group level could be affected by the nature of speech (i.e. expectations), there could be individual differences in the evaluations determined by listener-internal factors such as their personality traits (e.g. Dewaele and McCloskey 2015).

One of the current views of personality traits describes them as 'abstractions used to either explain or summarize coherent ABC [affective, behavioral, cognitive] (and sometimes D [desire]) patterns over time and space' (Ortony, Norman, and Revelle 2005; Wilt and Revelle 2015, 2). It is assumed that personality traits can account for variability in a broad range of specific behaviours (Condon et al. 2020; Matthews, Deary, and Whiteman 2009).

Several models and theoretical conceptions in the field of individual differences have supported the view that personality traits affect language evaluations. Among the early attempts, the Accent Prestige Theory (APT) proposed by Giles and associates in 1970s (Fuertes, Potere, and Ramirez 2002; Giles 1970; Giles and Powesland 1975) suggested two highest-order dimensions of interpersonal evaluation, Status and Solidarity, as the roof constructs to account for the variance in non-standard accent perception. While Status refers to social standing and position, Solidarity implies a thoughtful and friendly demeanour. The similarities are striking between the APT traits and two key constructs (i.e. Dominance and Affiliation) of the Interpersonal Circumplex, which is arguably the most influential interpersonal model in psychology of individual differences (Wiggins 1996). Status is a transparent counterpart of Dominance, and Solidarity of Affiliation. These similarities could be extended to the 'Big Five' personality model, given that Dominance and Affiliation are factor-rotated variations of Extraversion and Agreeableness in the Big Five (DeYoung et al. 2013; Wiggins 1996). The association suggests that these two personality traits are particularly relevant for interpersonal evaluations.

The Big Five (BF) model is lexical in nature and refers to five broad personality dimensions – Neuroticism/Emotional Stability, Extraversion, Intellect/Openness, Agreeableness, and Conscientiousness (De Raad and Mlačić 2020; John and Srivastava 1999). High scorers on Neuroticism tend to be 'worrying and emotional' due to 'psychological distress'; pronounced Extraversion suggests proneness to 'interpersonal interaction' and 'optimism'; Openness relates to appreciation of new experience and tolerance of unfamiliar things; higher scorers on Agreeableness are 'soft-hearted and helpful'; and high scorers on Conscientiousness are better at organisation, persistence and achieving goals (Cervone and Pervin 2013, 265). We opted for the Big Five model of personality as our conceptual framework, along with a conceptually close Five-Factor Model of Personality (FFM; Costa and McCrae 1985), as it is globally recognised as one of the staples of personality psychology, widely used in LX studies, and validated in a large number of studies, cultural and cross-cultural (Chen et al. 2022; De Raad and Mlačić 2020; John and Srivastava 1999).

In terms of language evaluations specifically, a few studies have reported substantial effects of personality traits on the ratings of foreign accent or grammatical errors. For example, L1 listeners who were extraverted or emotionally stable (low on Neuroticism) showed more tolerance to foreign accents, whereas more conscientious listeners rated the accent as stronger (Dewaele and McCloskey 2015; Gaffney and Côté 2019). Similarly, Boland and Queen (2016) found that grammatical errors in written texts were less penalised by L1 readers who were more extraverted or agreeable. However, the exact personality traits that showed effects vary across studies. This may be due to the potentially insufficient capacity of linear effects models to account for the link between personality and language evaluations. Indeed, several studies in personality research have demonstrated non-linear relationships among personality traits and various concepts of ability and behaviour, and reported more accurate predictions made by non-linear models compared with linear ones (Austin, Deary, and Gibson 1997; Ben-Ner and Kramer 2011; Bozionelos et al. 2014; Vasilopoulos, Cucina, and Hunter 2007). To address this issue, non-linear models will be used in this study to examine the effect of personality traits on LX speech evaluations.

Another methodological consideration of this study is the interaction between personality and linguistic conditions (i.e. accent and grammaticality), as the effect of personality could be subject to the experimental manipulation or situational context (e.g. Revelle 2007). According to Hubert Lyall and Järvikivi (2021), lower scorers of Openness showed greater difficulty (indicated by pupil dilation) in processing speech with grammatical errors, while less extraverted listeners showed greater difficulty in processing semantic anomalies and socio-cultural violation. Such 'contextualization' may be tentatively related to components of Funder's (2016) Situation Construal Model (SCM). To transpose SCM concepts to our study settings, verbal communication can be seen as a 'situation' variable, and speech evaluations as related to situation construal, connected to personality and situation by subjectivity (stemming from personality) and accuracy (reflected in ratings).

The current study

In summary, the current study attempts to answer whether and how L1 listeners' personality traits, speech accent and speech grammaticality interact to affect their evaluations of LX speech. More specifically, we examined the tolerance ratings of LX speech in different linguistic conditions (well-formed vs. error-filled, British-accented vs. Polish-accented) given by 60 L1 speakers of British English, and their personality traits were captured by the Big Five.

This study focused on English article errors which can occur frequently (Dryer 1989, 86) but are rarely corrected or eliminated. Article errors appear more tolerable than other types of grammatical errors (e.g. Ensz 1982; R. Johnson and Jenks 1994; Magnan 1982; Vann, Meyer, and Lorenz 1984), and are also very difficult to eradicate as article usage eludes accurate description (Yoon 1993; for a linguistic account of the challenges article usage poses, see Divjak, Romain, and Milin 2023). Therefore, errors against articles are very frequent in LX speech, which means that listeners may have habituated to and expect them in the context of a foreign accent.

Drawing on expectation mismatch effect (Gnevsheva 2018; McGowan 2015) and error detection research (Hanulíková et al. 2012), it is hypothesised that compared with foreign-accented speech, L1 listeners' ratings of native-accented speech would decline more when grammatical errors occur, due to unexpectedness (i.e. an interaction effect of speech accent and grammaticality). We also expect to find personality effects on LX speech ratings, which could be non-linear, and their interactions with experimental manipulation (grammaticality) and/or situational context (accent). Given that this study breaks new ground, we derive general hypotheses from the personality literature and expect the LX speech ratings to be more strongly related to those traits that are more saturated by interpersonal interaction (e.g. Extraversion and Agreeableness). Through estimation of effect sizes and interactions of different variables, we examine whether personality traits have a

general impact on behaviour (unaffected by linguistic conditions), or more context-specific effects (shaped by linguistic conditions).

Methods

Participants

Sixty L1 speakers of British English were recruited via social media (e.g. Facebook, Twitter) and among colleagues at a UK university. All of them were either born in the UK or had immigrated to the UK before they were four years old and spoke English as their primary language thereafter. The age of participants ranged widely from 18 to 71 ($M = 36.5$, $SD = 11.7$). No hearing difficulties were reported by participants or observed by the researcher during data collection. To generalise the findings to a wider population, we kept a balance in participants' gender (30 females vs. 30 males) and recruited participants with varied exposure to foreign-accented speech (with frequency of conversations with foreign-accented speakers of English ranging from daily, weekly, monthly, yearly, to never). Although participants reported varying familiarity with foreign-accented English speech on a 100-point rating scale ($M = 59.8$, $SD = 33.0$), after they completed the speech evaluation tasks, all but one thought that the foreign accent of the LX speech they heard could be Polish. Participants also showed variability in their education level – ranging from GCSE or lower to postgraduate degree level with 44 of them holding an undergraduate degree; language learning experience – 23 of them spoke at least one foreign language (mainly Romance languages); and English teaching experience – 12 of them had English teaching experience ranging from 1 to 35 years ($M = 9.2$, $SD = 9.9$).

Instruments

The participants' evaluations of LX English speech were measured via an auditory acceptability judgment task and their personality profiles were assessed through a short version of Big Five questionnaire (Donnellan et al. 2006). All tasks were administered via Qualtrics.

Auditory acceptability judgment task

The auditory acceptability judgment task contained 40 English speech stimuli (short passages) with or without grammatical errors and spoken in different accents. Participants' main task was to listen to each stimulus and rate the acceptability of it on a 100-point rating scale.

The text material of the speech stimuli was adapted from transcripts of the spoken interview provided in the BACKBONE English as Lingua Franca (ELF) Corpus produced by Polish speakers of English. Forty short passages extracted from eight individual interviews on different topics (e.g. education, environment, culture, globalisation, health, society, etc.) were used in this study. First, the 40 passages were checked carefully by a linguist who was an L1 speaker of British English, and all typos and grammatical errors were corrected. Then, different types of article errors were added to half of the passages (i.e. two to three passages randomly selected from each topic), including omission of definite and indefinite articles ('the' or 'a/an' → zero article), substitution of definite articles for indefinite ones or vice versa ('the' → 'a/an' or 'a/an' → 'the'), and overuse of articles (zero article → 'the' or 'a/an'). As a side-effect of using naturalistic stimuli, the density of article errors (i.e. the number of errors divided by the number of nouns) varied across those 20 passages from 18% to 56%, and the final length of the passages varied from 41 to 109 words ($M = 78$). The error condition of each stimulus (well-formed vs. error-filled) was used for the grammaticality measure as an independent variable.

Another independent variable, speech accent, was binary too (native vs. foreign). To create speech stimuli in native and foreign accents, each of the 40 passages was read out loud clearly by two British and two Polish speakers (one female and one male) at a natural speed and recorded

Table 1. Latin square for stimuli condition allocation.

Topics/ Blocks	Speakers			
	British female	Polish male	British male	Polish female
1	A	B	C	D
2	D	A	B	C
3	C	D	A	B
4	B	C	D	A
5	A	B	C	D
6	D	A	B	C
7	C	D	A	B
8	B	C	D	A

using a phone in a quiet room. The length of speech stimuli varies from 13 to 39 s ($M = 25$). Then, we used a Latin square design to allocate all 40 passages in different speaker voices to each participant (as shown in Table 1). The four sets of speech stimuli (A, B, C, D) were allocated to participants randomly. Each five speech stimuli were about the same topic and recorded by the same speaker, and they were presented to participants as a block (i.e. in a fixed order), with the order of blocks varying for the four sets – 1-2-3-4-5-6-7-8 for set A, 3-4-5-6-7-8-1-2 for set B, 8-7-6-5-4-3-2-1 for set C, and 4-3-2-1-8-7-6-5 for set D.

The acceptability rating of each speech stimulus was elicited as an overall perception of the naturalness of LX speech, via a question 'How does the language in this recording sound to you?', with two anchor points on the 100-point rating scale (1 = This is not how I would express myself in English; 100 = This is pretty much how I would put it myself). These statements were used as a way of probing the effect of grammatical errors and foreign accents without drawing attention to them, while being specific enough to elicit clear opinions (vs. a label of 'acceptable' or 'unacceptable'). As the acceptability ratings could be affected by the difficulty in understanding the speech due to foreign accents or grammatical errors, participants were asked to assess the comprehensibility (i.e. perception of the ease of understanding; Munro and Derwing 1995) of each speaker after listening to every five speech stimuli, by answering questions 'How easy was it to understand the speaker?' (1 = very difficult, 100 = very easy). As shown in Table 2, most LX speech stimuli were rated as highly comprehensible, with over 90% of the comprehensibility scores above 70.

The speech stimuli were presented to participants at a rate of one stimulus per page on Qualtrics. JavaScript was used to make sure that participants played each stimulus only once before giving an acceptability rating. The audio player was hidden immediately after the 'Play' button was pressed, so it could not be pressed again. The 'Next' button was hidden when the page loaded. Only after the stimulus was played and a rating score was given, could participants click the 'Next' button to proceed to the next stimulus.

Table 2. Rating score distribution of comprehensibility of stimuli.

Score (1–100)	Comprehensibility
> = 95	56.9%
> = 90	10.4%
> = 85	8.1%
> = 80	6.5%
> = 75	5.2%
> = 70	4.4%
<70	8.5%
> = 65	2.5%
> = 60	1.3%
> = 55	1.5%
> = 50	0%
<50	3.3%

Personality questionnaire

The Big Five model of personality was measured by the 20-item Mini-IPIP scale (Donnellan et al. 2006), adapted from the 50-item International Personality Item Pool-Five Factor Model (IPIP-FFM; Goldberg 1999). It has been validated as a short measure of the Big Five factors of personality, with four items allocated to each factor – Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Intellect/Imagination (or Openness). Participants gave responses on a 5-point Likert scale from 1 = 'completely disagree' to 5 = 'completely agree' with 'neither agree nor disagree' in the middle. Some of the items were reverse coded (see Appendix A for the questionnaire). Cronbach alpha coefficients (calculated using psych package in R: Revelle 2021) pointed to acceptable reliability (Extraversion: α = .83; Agreeableness: α = .83; Conscientiousness: α = .72; Neuroticism/Emotional Stability: α = .73; Intellect/Imagination/Openness: α = .82).

To get a composite score for each personality trait while taking the weights of each item into consideration, a set of principal components analyses was run on questionnaire items measuring each trait, using the princomp () function in the R statistical environment (R Core Team 2022). The first principal components, with the most balanced loadings of items, were extracted for each trait, and the scores were used in subsequent analyses.

Questionnaire for other personal information

Participants also reported their demographic background and language-related experience and attitudes via a questionnaire, including age, gender, education level, language learning (the number of languages they spoke other than English) and teaching experience (whether they had taught English), linguistic training experience (whether they had studied linguistics), frequency of exposure to foreign-accented speech (daily, weekly, monthly, yearly, or never), and familiarity with and attitudes to foreign accents. These personal variables were used as control variables in the statistical analyses as they may affect L1 listeners' evaluations of LX speech.

Certain questions were asked during recruitment and at the beginning of the experiment to guarantee a balance in gender (30 female vs. 30 male) and varied exposure to foreign-accented speech, i.e. 30 participants conversed with foreign-accented speakers weekly or more often (N_{daily} = 19, N_{weekly} = 11) and the other 30 did so monthly or less often ($N_{monthly}$ = 11, N_{yearly} = 15, N_{never} = 4). This variability mapped onto the variation in participants' familiarity with foreign accents (M = 60, range: 30–100), self-reported via the question 'How familiar are you with foreign-accented English speech' (1 = 'not at all, 100 = 'very much').

As for participants' attitudes toward foreign accents, they were asked to choose from a list of adjectives to describe accented speech and/or to provide any other adjectives not on the list. The adjectives represented an emotion continuum, ranging from 'attractive', 'exciting', 'romantic', 'interesting', 'pleasant', 'funny' to 'unpleasant', 'distracting', 'annoying', 'intimidating', and 'terrifying'. In Warriner, Kuperman, and Brysbaert (2013), about 14,000 English lemmas were rated for their emotional valence on a 9-point rating scale (1 = 'completely happy' and 9 = 'completely unhappy'), including the adjectives used in our study. To calculate a composite score, the emotional valence score of each adjective was firstly recentered to 0 (range: – 4–4) to reflect positive and negative emotion, and then the scores of all adjectives chosen and/or added by each participant were averaged. All participants but two got positive scores (M = 1.65, range: – 1.32–2.60), which suggests that almost all of them showed positive attitudes to foreign accents. Note that it is not a result of restricted recruitment as there was no mention of 'attitudes to accent' in the recruitment message. Rather, our sample is representative of British speakers who live in a big city in the UK.

Procedure

Due to the lockdown triggered by the COVID-19 pandemic, the data collection had to be done online. To control the quality of the data, all tasks and questionnaires were administered via individual video meetings between participants and the researcher, who is a highly proficient L2 English

speaker of Asian origin. Participants were asked to use a computer and a set of headphones in a quiet room. A session took around 45 min, including a 5-minute introduction, an auditory acceptability judgment task (25 min) and questionnaires for personality and other personal information (15 min). After a brief introduction of the procedure, a link to the auditory acceptability judgment task was sent to participants. When they finished the first part, a link to the questionnaires was sent to them. During the tasks, participants were asked to leave the camera on, whereas the researcher turned off the microphone and camera to avoid disturbing them. Upon the completion of all tasks, each participant received a £7.5 Amazon voucher as reimbursement.

Power analysis

Power analysis was run via a prototype model (Wang and Rhemtulla 2021), which included the linear and non-linear effects of null, small, medium and large sizes. It confirmed that the sample size was sufficient to ensure acceptable statistical power in detecting effects of all sizes, and dismissing non-existent (null) effects, regardless of their linearity (See Appendix B for details).

Results

A series of data analyses was conducted to examine the potential interaction effects of speech accent, speech grammaticality and listener personality traits on the acceptability ratings of LX speech, using the R software environment (R Core Team 2022). Each of the 60 participants gave an acceptability rating score to each of the 40 speech stimuli, yielding a total of 2,400 data points. The dependent variable, Acceptability (rating scores for the speech stimuli), was predicted from linguistic conditions of Accent (British vs. Polish) and Grammaticality (well-formed vs. error-filled), and listener personality traits (Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness), while controlling for speech Comprehensibility (max vs. sub-max scores) and Speaker Gender, along with listeners' Age, Gender, Education Level, experience of Language Learning, English Teaching, Linguistic Training, Frequency of Exposure to foreign-accented speech, Familiarity with Foreign Accent, and Attitudes to Foreign Accent.

Around 22% of rating scores for British-accented stimuli and 12% of rating scores for Polish-accented stimuli obtained the maximum score of 100. Removing the maximum scores left us with 83% of the total data (1995 datapoints). See Appendix C for more details about data distribution and the justification for removing the maximum acceptability scores. The remaining datapoints were rank-to-normal transformed to facilitate statistical modelling (c.f., Aulchenko et al. 2007; N. L. Johnson 1949) and a further 24 extreme datapoints (> 3 standard deviations) were visually identified and trimmed. This resulted in the final dataset with 1971 datapoints in total (82% of the original data).

Generalized Additive Mixed-Effects Models (GAMMs) were fitted to the non-max Acceptability ratings scores using the mgcv and itsadug packages in R (version 1.8-33 for mgcv, Wood 2011, 2017; version 2.3 for itsadug, van Rij et al. 2017). A GAMM can be understood as a superclass of many related models, including ANOVA/ANCOVA and Linear and Generalized Linear Regression, including mixed-effects models. It allows capturing non-linear relationships between a particular covariate (numeric predictor) and the dependent variable, alongside inclusion of random effects for participants and items. In our case, all models were ANCOVA-like, as we tested both categorical factors (e.g. Accent, Grammaticality) and continuous covariates (e.g. personality scores); in the case of the latter, we examined if their relationship with the Acceptability rating scores was non-linear. GAMMs are prone to overfitting if nonlinearity is uncontrolled. Our approach was rather conservative as we allowed only minimal non-linear smooths for straightforward interpretation. Hence, the maximum number of knots was set to 4 for all fixed effect smooths.

A candidate model, once determined, should be checked to see whether its effects are driven by influential values. Somewhat simplified, we can distinguish two types of influential values: extreme

Table 3. Outputs of the generalised additive mixed-effects model of the association between acceptability scores and predictors: ML = −1791.5; R-sq. (adj) = 0.671; Deviance explained = 70.9%.

| A. Parametric coefficients	Estimate	Std. error	t-value	p-value
Intercept	0.850	0.095	8.952	< .0001
Speaker Gender (Male)	−0.220	0.030	−7.336	< .0001
Comprehensibility (Submax)	−0.359	0.047	−7.655	< .0001
Grammaticality (error-filled) * Accent (Polish)	−0.648	0.039	−16.804	< .0001
B. Smooth terms	edf	Ref.df	F-value	p-value
Conscientiousness * Accent (British)	1.000	1.000	0.852	0.356
Conscientiousness * Accent (Polish)	1.000	1.000	4.806	0.029
Extraversion * Accent (British)	1.000	1.000	0.722	0.396
Extraversion * Accent (Polish)	2.896	2.981	16.057	< .0001
Agreeableness * Accent (British)	1.000	1.000	0.457	0.499
Agreeableness * Accent (Polish)	2.376	2.685	5.061	0.016
Trial Order * Participant ID	200.880	533.000	4.495	< .0001
Block	6.033	7.000	8.322	< .0001

values (which *inflate* an effect) and outliers (which *deflate* an effect). Following Baayen and Milin (2010), we trimmed extreme residuals (> 2.5 standard deviations from the mean; 99% interval) and re-fitted the model to confirm that there were no effects due to influential values. A total of 42 extreme residuals were removed and all significant effects remained significant. Finally, we also included random effects for both participants and items (for applications of these statistical techniques to language data; see Divjak, Milin, and Medimorec 2020; Milin et al. 2017; for more in-depth discussion, see Baayen et al. 2017; Matuschek et al. 2017; Wood 2017).

The final, best model (as shown in Table 3) includes the main effects of Speaker Gender and Comprehensibility and an interaction of Grammaticality by Accent. Accent also entered into interactions with three measures of personality traits: Conscientiousness, Extraversion, and Agreeableness. As random effects, the final model contained intercept adjustments for experimental Block (each containing five stimuli about the same topic) as well as factor smooths of Trial Order (the presentation order of each stimulus, z-transformed) by Participant ID. Other variables submitted to the model did not show any significant effects on Acceptability. As summarised in Table 3, the model has 71% of explained deviance (with).

Figure 1 depicts the parametric terms. All Y-axes reflect scaled acceptability scores with values at the bottom indicating lower Acceptability and values at the top indicating higher Acceptability. According to the left panel, compared to stimuli recorded by female speakers, stimuli recorded by $mR^2_{Adj.} = 0.67$ale speakers received lower acceptability ratings. The middle panel shows that when the comprehensibility of the stimuli was not rated as perfect, the acceptability rating was likewise reduced. In the right, the significant interaction of Grammaticality by Accent indicates that,
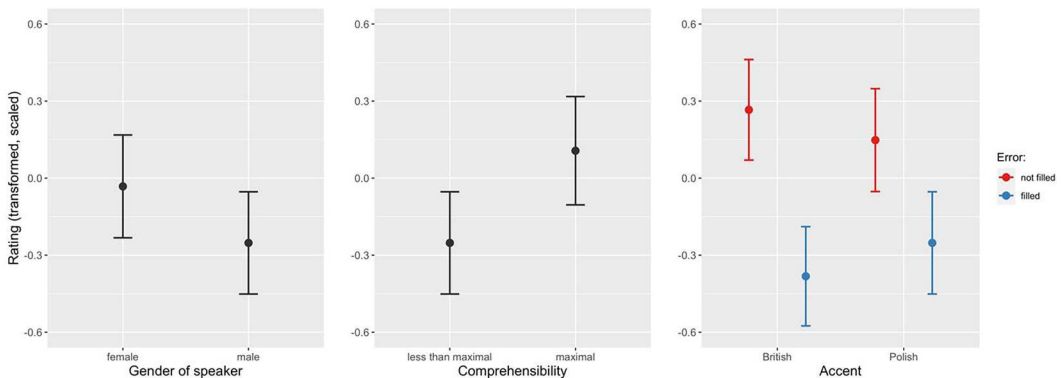
Figure 1. Effects of speaker variables on acceptability scores.
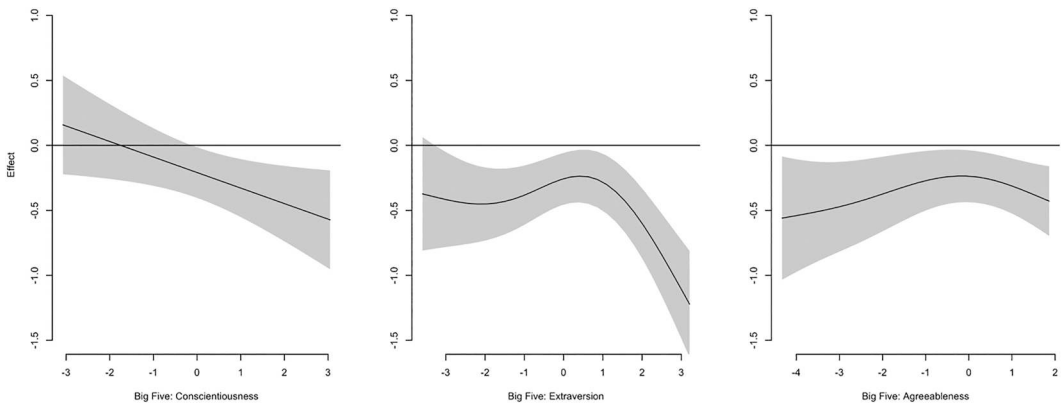
Figure 2. Effects of listener personality on acceptability scores (in Polish accent condition).

the occurrence of article errors in British-accented speech was more detrimental than that in Polish-accented speech. In other words, article errors appeared to be more acceptable if they were made by a foreign-accented speaker.

For the interpretation of the smooth terms, we rely on their visual summaries in Figure 2. Significant effects of three personality traits (i.e. Conscientiousness, Extraversion, Agreeableness) were found only for Polish-accented speech. In all plots, across the X-axes, the zero indicates the midpoint, with values to the left indicating increasing degrees of absence of the trait, while values to the right indicate increasing degrees of presence of a trait. Conscientiousness, on the left panel, had a linear effect (edf = 1) on acceptability ratings such that listeners who were more conscientious assigned lower scores to Polish-accented speech. The effects of both Extraversion (mid panel) and Agreeableness (right panel) on the ratings were non-linear, elbow-shaped. It shows that Extraversion had a negative effect among listeners with above-average trait scores, i.e. more extraverted individuals gave lower acceptability ratings. The trend is, however, not as obvious among listeners with below-average trait scores – there seems to be a positive effect of Extraversion from – 1.5–0 on the scale, but it is rather moderate. The effect of Agreeableness is moderate overall, with a positive correlation with ratings for the range of values between – 4 and 0, such that more agreeable listeners appear to give higher acceptability scores; after that, the trend reverses to a negative correlation for values higher than 0, suggesting that more agreeable listeners gave lower ratings. These patterns will be further discussed in relation to the literature in section 4.2.

Discussion

This study investigated how L1 listeners' personality traits and the accent and grammaticality of LX speech interact to determine language evaluations. We achieved this by examining the acceptability ratings assigned by L1 listeners to English speech stimuli, with the accent (British vs. Polish) and grammaticality (well-formed vs. filled with English article errors) systematically manipulated; listeners' personality was assessed using a standard personality measure (Big Five).

It is found that the presence of grammatical errors in native-accented speech caused a larger decline in acceptability ratings than they did in foreign-accented speech, possibly because grammatical errors in native-accented speech were more unexpected and thus more negatively perceived by L1 listeners. On the other hand, listeners showed individual differences in their ratings. Certain personality traits (Conscientiousness, Agreeableness, Extraversion) affected how they rated foreign-accented speech, regardless of the presence of grammatical errors. In terms of the controlled variables, speaker gender and comprehensibility showed significant effects on acceptability ratings – when listening to speech produced by male speakers or speech that was not rated as 100%

comprehensible, L1 listeners rated the speech as less acceptable. Other listener variables (e.g. language-related attitudinal and experiential factors), however, showed neither significant effect on acceptability ratings nor interaction with accent or grammaticality.

The role of listener expectations

The interaction effect of accent and grammaticality suggests a general pattern that L1 listeners, at the group level, reacted more negatively to grammatical errors in native-accented speech than to errors in foreign-accented speech. The result echoes recent findings about L1 readers' perception of grammatical errors (Kirjavainen et al. 2024) and supports the expectation mismatch effect (McGowan 2015).

Research on memory shows that when the probabilities of co-occurrence between two things are high, the combination is more likely to be stored in memory and will hence be expected to occur (Divjak 2019). Although participants' frequency of exposure to foreign-accented speech varied widely, only four of them reported no exposure at all. This suggests that most of them had been exposed to article errors, one of the most frequent errors made by LX speakers, in foreign-accented English speech. Thus, they had likely habituated to article errors and were not surprised to hear them in a foreign accent. In contrast, article errors occur very rarely in L1 speech and thus, listeners did not expect these errors to occur in native-accented speech and tended to notice them more. As a result, they rated error-filled, native-accented stimuli as less acceptable (i.e. expectation mismatch effect). This interpretation is in line with how Hanulíková et al. (2012) explained their findings that a P600 effect was detected in L1 listeners' online processing of gender agreement errors in native – but not foreign-accented Dutch speech.

The role of personality traits

At the individual level, three personality traits (Conscientiousness, Agreeableness, Extraversion) predicted L1 listeners' ratings of foreign-accented speech. The results confirm our hypotheses that personality traits, especially those closely associated with interpersonal interaction, have an influence on LX speech evaluations, and that the personality effects are moderated by the situational context of the evaluations (i.e. speech accents). What is more, as we expected, the effects of personality traits (specifically Extraversion and Agreeableness) are non-linear. We will discuss personality effects in relation to each of the three traits in the light of their (linear or non-linear) interactions with speech accent.

Considering the fundamentally interpersonal nature of the context in which the ratings were provided, one could argue that the predictive power of Extraversion and Agreeableness, two personality traits that are most involved in communication and interpersonal processes, speaks in favour of the validity of the Interpersonal Circumplex model (DeYoung et al. 2013; Wiggins 1996), given the associations between the key constructs of the model (Dominance and Affiliation) and these two traits. More importantly, Extraversion and Agreeableness demonstrated complex and non-linear effects on the speech evaluations.

A moderate positive effect of both personality traits appeared among L1 listeners who scored (just) below average on the scales. That is, the more (moderately) extraverted or agreeable they were, the more acceptable foreign-accented speech sounded to them. This trend is consistent with previous findings where L1 listeners who were more extraverted or agreeable (with marginal significance) rated foreign-accented speech more leniently (Gaffney and Côté 2019), and L1 users who were more extraverted were slightly less bothered by foreign accents (Dewaele and McCloskey 2015). However, among listeners who scored above average on Extraversion or Agreeableness, the trend reversed. This pattern was more pronounced for the strongest Extraverts and only mild for the most Agreeable participants: the more (extremely) extraverted or agreeable they were, the lower the acceptability ratings they gave to foreign-accented speech. Although the negative effects of

Extraversion and Agreeableness on LX speech evaluations seem contrary to the above-mentioned previous findings (Dewaele and McCloskey 2015; Gaffney and Côté 2019), similar results have been reported in recent studies on rating severity. For example, more extraverted LX raters were found to be harsher in their ratings of LX writings (Choi and Lee 2019; Zhu, Fung, and Yang 2021), and more agreeable raters were found to give lower scores in image ratings (i.e. the degree they liked an image; Aniceto, Gena, and Vernero 2023). The non-linear effect of personality traits has been confirmed repeatedly in studies on various types of behaviour. For instance, Curşeu et al. (2019) found and replicated substantial quadratic relations between Extraversion, Conscientiousness and contributions to teamwork. Bozionelos et al. (2014) pointed to the relevance of non-linear relations among personality traits and the perception of professional mentoring received, with consistent quadratic effects of Agreeableness and Conscientiousness across two studies.

While our results revealed non-linear effects of Extraversion and Agreeableness, suggesting that personality effects could be modulated by the zones on the trait continuum raters are mapped onto, we cannot pinpoint the exact source of these 'paradoxical' effects. However, we attempt to offer some (speculative) explanations, focusing more on the negative effects of the two traits as they appear to be counter-intuitive and contrary to most previous findings. As extraverts tend to be more 'sociable and optimistic' and agreeable individuals tend to be more 'soft-hearted and forgiving' (Cervone and Pervin 2013, 265), they are more likely to be tolerant to foreign-accented speakers (e.g. Gaffney and Côté 2019), hence the positive effects. This is in line with the interpretation in previous studies (Dewaele and McCloskey 2015; Gaffney and Côté 2019). So, why did the pattern reverse in our study among L1 listeners scoring high on these traits?

As for Extraversion, the literature points to 'classic' theories of Extraversion, such as arousal theory: 'extraverts should outperform introverts in highly arousing situations' (Wilt and Revelle 2009, 32; Yerkes and Dodson 1908), and that 'extraverts should, on average, respond more and faster than introverts' (Wilt and Revelle 2009, 32; Wundt and Judd 1897). Hence, one may assume that the reversed effect in extreme scores on Extraversion may have occurred due to extreme extraverts' expectedly faster and less careful response pattern in a 'repetitive' experimental task. Such pattern may be linked to the impulsiveness trait, which, though not an integral part of Extraversion, has been linked to it in several models (Revelle 1997). This interpretation is to some degree supported by findings in Cokely and Feltz (2009) who found that extraverts were more likely to show affective bias possibly due to their 'looser regulation of affective reactions' (19).

Indeed, one facet of Extraversion, Assertiveness, seems to echo the assumption made above. Facets refer to more specific personality features, or lower-order traits, underlying the upper-level personality trait. Cervone and Pervin (2013, 272) have summarised six facets underlying Extraversion, including Gregariousness, Activity Level, Assertiveness, Excitement Seeking, Positive Emotions, and Warmth. Compared with introverts who are usually 'reserved' and 'aloof', extraverts may be more likely to express their opinions freely (Assertiveness), due to the higher level of interaction and activity (Cervone and Pervin 2013, 265). Thus, when asked about the degree to which the foreign-accented speech stimuli could represent them, it may have been easier for more extraverted listeners to point out the gap between the stimuli and the standard norm (i.e. native accent), yielding lower ratings. Assuming that individuals with extreme scores on Extraversion are more sociable, communicative, and talkative, they may also be more susceptible to speech stimuli and more eager to react to potential mistakes (cf. Choi and Lee 2019). Thus, it could be the case that on the positive side of the Extraversion scale, the facet Assertiveness showed a stronger effect than the facet Positive Emotions.

In a similar vein, among the underlying facets of Agreeableness (Straightforwardness, Trust, Altruism, Modesty, Tendermindedness, Compliance; Cervone and Pervin 2013, 272), Straightforwardness may have had a bigger impact on individuals scoring highly on Agreeableness than Tendermindedness. As a result, these listeners tended to express their opinions honestly and directly, hence the lower acceptability ratings. Further investigation is required with a rater sample of greater variability in each personality trait and a more 'finely tuned' conceptualisation of the traits, i.e. a

more extensive personality measurement that differentiates between the facets (e.g. BFAS, DeYoung, Quilty, and Peterson 2007; AB5C; Hofstee, De Raad, and Goldberg 1992; Big Five 2; Soto and John 2017). Additionally, future studies could use interviews, stimulated recall or think-aloud methods to investigate the mental processes involved in the rating (cf. Bowles 2010; Gass and Mackey 2017; Padilla and Leighton 2017).

Finally, in terms of Conscientiousness, results showed that more conscientious listeners penalised foreign-accented speech more. Adherence to rules and principles, generally attributed to conscientiousness, may be the first step towards understanding this negative effect. As conscientiousness individuals tend to be 'ambitious, hard-working and self-disciplined' (Cervone and Pervin 2013, 265) according to personality descriptors in the AB5C model (Hofstee, De Raad, and Goldberg 1992), they may be motivated to speak a language in the most standard way. Indeed, Conscientiousness is one of the traits 'most commonly associated with exceptional L2 attainment' (Moyer 2021, 64). Thus, when foreign-accented speakers could not meet the standard norm (i.e. native accent), it seems reasonable that listeners who were adhering to the rules more strictly would find this less acceptable. The result is in line with findings in Gaffney and Côté (2019) where more conscientious listeners tended to rate LX speech as more accented.

The interactions between personality traits and accent conditions suggest that L1 listeners responded to foreign-accented speech differently depending on their personality traits (regardless of grammaticality), but such personality effects did not spill over to their ratings of native-accented speech. It may help to consider associations between personality and potential prejudice towards outgroups for which a foreign accent is a powerful cue (Gluszek and Dovidio 2010). For instance, a large-scale study conducted by Lin and Alvarez (2020) pointed to direct associations between Agreeableness and Conscientiousness with anti-black prejudice, while controlling for social and ideological attitudes (authoritarianism, social dominance orientation, and party affiliation). Thus, personality traits could contribute to the formation of judgments and expressions of attitudes that may be vital for establishing and sustaining various forms of prejudice, and the prejudice could then be reflected in LX evaluations. The complexity of such associations, however, calls for replication and further investigation.

As no interaction was found between personality and grammaticality, L1 listeners with different personality profiles in our study did not react to grammatical errors differently, whereas in Boland and Queen (2016), less agreeable and less extraverted L1 readers were more sensitive to grammatical errors in text. This could be due to the fact that accent condition was not included in Boland and Queen (2016). Although both foreign accents and grammatical errors point to language deviations from the standard norm, a foreign accent is a more salient feature than grammatical errors, so when both conditions exist, listeners would respond more to accent. The result suggests that personality traits may display the same behaviour but for different reasons.

In summary, although the effects of some traits were moderate, the non-linear interaction effects of personality and accent support the importance of taking contextual factors and nonlinearity into consideration in the construction of personality models (e.g. Situation Construal Model: Funder 2016). Further studies would be necessary to demonstrate the replicability of such results, and possibly to take it further by including the factor-rotated versions of Extraversion and Agreeableness in the circumplex model (Dominance and Affiliation). Mostly due to the sample size, we did not address the issue in this study.

As L1 listeners' evaluations of foreign-accented speech could have negative consequences, especially in high-stakes situations (e.g. academic admissions, job interviews, immigration interviews), it is important to raise their awareness of the role their personality traits could play in ratings, which may help the listeners/raters in the communication with and appreciation of LX speakers. In fact, as pointed out in Derwing and Munro's (2014) review, listeners can improve their perception of LX speech effectively and efficiently via training that stimulates perceptual adaptation, such as awareness-raising activities that facilitate the listeners' understanding of the potential biases against foreign accents, contact activities which help them notice the differences and

similarities between L1 and LX speech, and linguistic training that enhances their comprehension of foreign-accented speech (see Miao, Moran, and Kang 2023 for the effectiveness of awareness training; cf. Roessel, Schoel, and Stahlberg 2020).

The current study contributes to the development of training approaches by revealing that the role of listeners' personality traits in their ratings is not a simple linear effect, and that they are more likely to respond to foreign accents in LX speech evaluations than to grammatical errors. Further research is needed before our findings can be implemented in real-life settings.

The role of controlled variables

In addition to the interaction effects of accent by grammaticality and accent by three listeners' personality traits, we also found the main effects of two controlled variables (i.e. Comprehensibility and Speaker Gender).

Firstly, the speech stimuli rated as maximally comprehensible (i.e. scored 100 on a 100-point scale of comprehensibility) were perceived as more acceptable than the rest of stimuli. The result is not surprising as more comprehensible speech requires less time to process (Munro and Derwing 1995). More importantly, it suggests that the effects of accent, grammaticality and listener personality are independent of the effect of the speaker's comprehensibility.

Secondly, L1 listeners evaluated LX speech produced by female speakers as more acceptable than that produced by male speakers. Since we carefully controlled the balance in the gender of both listeners and speakers (i.e. each of the 30 female and 30 male listeners rated 20 speech stimuli in female voices and 20 in male voices) and there was no interaction between speaker gender and listener gender, the effect of speaker gender on acceptability cannot be attributed to the potentially (dis)favoured perception of the same/different gender.

It is important to note that the generalizability of the finding should be approached with caution, given that only one speaker was used for each of the four accent-gender experimental conditions (female or male with British or Polish accent). Introducing more speakers for each condition would have introduced uncontrolled variability, which we aimed to avoid. Unarguably, however, similar to accent, the gender of the speakers was one of the most salient features of the speech samples, overshadowing other potential voice characteristics. For example, female speakers were perceived to be more attractive than male speakers in Linek, Gerjets, and Scheiter (2010), which could be due to voice qualities such as stronger breathiness, greater girlishness or less creakiness (Babel, McGuire, and King 2014; Klatt and Klatt 1990; Levitt and Lucas 2018).

As for listener variables (e.g. language-related attitudes and experience), none of them showed any effects on L1 evaluations of LX speech. Based on previous findings, these variables should be either mediated by attitudes to foreign accent or via interactions with comprehensibility (e.g. Dewaele and McCloskey 2015; Kang and Rubin 2009). However, in our study, L1 listeners' reaction to accent and grammaticality was not mediated by their attitudes to foreign accent, and more language experience did not improve the comprehensibility of speech stimuli. The results could be due to the fact that participants' attitudes toward foreign accents were positive in general. Future studies may address this issue by recruiting participants with a greater variability in their language attitudes.

The use of a subjective measure for the acceptability judgment of LX speech would also benefit from further research. Instead of asking participants to rate how acceptable each speech stimulus was (with a scale from not acceptable to fully acceptable), we asked them to evaluate how the language in each stimulus sounded to them (with anchor points 1 = This is not how I would express myself in English, 100 = This is pretty much how I would put it myself). While this avoids problems with a subjective interpretation of 'acceptable', which may differ from rater to rater, it may make some readers wonder whether this measure taps into L1 listeners' own language use. Since L1 users rarely make English article errors, if our raters were reflecting on their own use of English, they should have given low ratings to all stimuli with article errors. However, our data showed

that the ratings they gave had great variability and even skewed towards max scores, which means that the ratings did indicate the evaluations of LX speech. Although the validity of this measure is supported by the data to some degree, further evidence would need to be collected from the participants which could stem from exit interviews, stimulated recall or think-aloud methods.

Conclusion

This study examined the degree to which L1 listeners' evaluations of LX speech could be explained by the interactions of speaker accent, grammaticality and listener personality. The interaction effect of accent and grammaticality further reveals the role of listener expectations in LX evaluations and supports the expectation mismatch effect. As the occurrence of grammatical errors in native-accented (vs. foreign-accented) speech was less expected, it was perceived more negatively by listeners, regardless of how comprehensible the speech was. On the other hand, the effect of listener personality was modulated by speech accent but not grammaticality, in that the traits of Conscientiousness and high levels of Extraversion showed negative effects, while Agreeableness showed more complex and mild effects on the ratings of foreign-accented speech, regardless of the presence of grammatical errors. Overall, the results suggest that foreign accent could 'protect' LX speakers from being penalised for grammatical errors to some degree due to listener expectations, unless the L1 listeners are highly conscientious or highly extraverted.

Theoretically and methodologically, our results reveal several points of interest for future studies, specifically concerning personality structure, the traits' interactions with contextual factors, and the specific mechanisms of traits' effects on observable behaviour. Namely, two out of three significantly predictive traits (i.e. Extraversion and Agreeableness) in our study have traditionally been highlighted as the most relevant driving forces of interpersonal evaluations, which points to further investigation of the relevance of the Interpersonal Circumplex model in LX research. Additionally, we have demonstrated the importance of examining personality effects while considering its interaction with contextual features and the more complex, i.e. non-linear relationships.

Note

1. We use LX as an umbrella term for any language acquired after the first language (L1), instead of L2 which has been used conventionally but may cause confusion when referring to L3, L4, etc. (Dewaele 2018).

Acknowledgements

Disclosure statement

Funding

Data accessibility statement

Data and code supporting this paper are available.
Code can be downloaded from https://github.com/ooominds/EvaluationOfLXSpeech and data is available at https://edata.bham.ac.uk/1181/.

Ethics approval statement

This study was approved by University of Birmingham Research Ethics Committee.

References

Aniceto, A., C. Gena, and F. Vernero. 2023, September. "The Influence of User Personality and Rating Scale Features on Rating Behaviour: An Empirical Study." In *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*, edited by C. Gena, L. De Russis, D. Spano, R. Lanzilotti, T. Di Mascio, C. Prandi, and S. Andolina, 1–8. New York: Association for Computing Machinery.

Aulchenko, Y. S., S. Ripke, A. Isaacs, and C. M. Van Duijn. 2007. "GenABEL: An R Library for Genome-Wide Association Analysis." *Bioinformatics (Oxford, England)* 23 (10): 1294–1296. https://doi.org/10.1093/bioinformatics/btm108.

Austin, E. J., I. J. Deary, and G. J. Gibson. 1997. "Relationships Between Ability and Personality: Three Hypotheses Tested." *Intelligence* 25 (1): 49–70. https://doi.org/10.1016/S0160-2896(97)90007-6.

Baayen, H., and P. Milin. 2010. "Analyzing Reaction Times." *International Journal of Psychological Research* 3 (2): 12–28. https://doi.org/10.21500/20112084.807.

Baayen, H., S. Vasishth, R. Kliegl, and D. Bates. 2017. "The Cave of Shadows: Addressing the Human Factor with Generalized Additive Mixed Models." *Journal of Memory and Language* 94:206–234. https://doi.org/10.1016/j.jml.2016.11.006.

Babel, M., G. McGuire, and J. King. 2014. "Towards a More Nuanced View of Vocal Attractiveness." *PLoS One* 9 (2): e88616. https://doi.org/10.1371/journal.pone.0088616.

Baese-Berk, M. M., D. J. McLaughlin, and K. B. McGowan. 2020. "Perception of non-Native Speech." *Language and Linguistics Compass* 14 (7): e12375. https://doi.org/10.1111/lnc3.12375.

Ben-Ner, A., and A. Kramer. 2011. "Personality and Altruism in the Dictator Game: Relationship to Giving to kin, Collaborators, Competitors, and Neutrals." *Personality and Individual Differences* 51 (3): 216–221. https://doi.org/10.1016/j.paid.2010.04.024.

Boland, J. E., and R. Queen. 2016. "If You're House is Still Available, Send me an Email: Personality Influences Reactions to Written Errors in Email Messages." *PLoS One* 11 (3): e0149885. https://doi.org/10.1371/journal.pone.0149885.

Bowles, M. A. 2010. *The Think-Aloud Controversy in Second Language Research*. London: Routledge.

Bozionelos, N., G. Bozionelos, P. Polychroniou, and K. Kostopoulos. 2014. "Mentoring Receipt and Personality: Evidence for non-Linear Relationships." *Journal of Business Research* 67 (2): 171–181. https://doi.org/10.1016/j.jbusres.2012.10.007.

Cervone, D., and L. A. Pervin. 2013. *Personality: Theory and Research*. Hoboken, NJ: John Wiley & Sons.

Chen, X., J. He, E. Swanson, Z. Cai, and X. Fan. 2022. "Big Five Personality Traits and Second Language Learning: A Meta-Analysis of 40 Years' Research." *Educational Psychology Review* 34 (2): 851–887. https://doi.org/10.1007/s10648-021-09641-6.

Choi, Y. H., and S. Lee. 2019. "The Role of Raters' Personality Traits in the Rating Process of L2 Argumentative Writing: Interactions with Argument Positions." *Journal of Research in Curriculum Instruction* 23 (2): 69–82. https://doi.org/10.24231/rici.2019.23.2.69.

Cokely, E. T., and A. Feltz. 2009. "Individual Differences, Judgment Biases, and Theory-of-Mind: Deconstructing the Intentional Action Side Effect Asymmetry." *Journal of Research in Personality* 43 (1): 18–24. https://doi.org/10.1016/j.jrp.2008.10.007.

Condon, D. M., D. Wood, R. Mõttus, T. Booth, G. Costantini, S. Greiff, W. Revelle, et al. 2020. "Bottom up Construction of a Personality Taxonomy." *European Journal of Psychological Assessment* 36 (6): 923–934. https://doi.org/10.1027/1015-5759/a000626.

Conover, W. J., M. E. Johnson, and M. M. Johnson. 1981. "A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data." *Technometrics* 23 (4): 351–361. https://doi.org/10.1080/00401706.1981.10487680.

Costa, P. T., and R. R. McCrae. 1985. *The NEO Personality Inventory Manual*. Odessa, FL: Psychological Assessment Resources.

Curşeu, P. L., R. Ilies, D. Vîrgă, L. Maricuţoiu, and F. A. Sava. 2019. "Personality Characteristics That are Valued in Teams: Not Always "More is Better"?" *International Journal of Psychology* 54 (5): 638–649. https://doi.org/10.1002/ijop.12511.

De Raad, B., and B. Mlačić. 2020. "The Big Five Personality Trait Factors." In *Oxford Research Encyclopedia of Education*. https://oxfordre.com/education/view/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-894

Derwing, T. M., and M. J. Munro. 2014. "Training Native Speakers to Listen to L2 Speech." In *Social Dynamics in Second Language Accent*, edited by J. M. Levis, and A. Moyer, 219–236. Berlin: De Gruyter.

Dewaele, J.-M. 2018. "Why the Dichotomy 'L1 Versus LX User' is Better Than 'Native Versus non-Native Speaker'." *Applied Linguistics* 39 (2): 236–240.

Dewaele, J.-M., and J. McCloskey. 2015. "Attitudes Towards Foreign Accents among Adult Multilingual Language Users." *Journal of Multilingual and Multicultural Development* 36 (3): 221–238. https://doi.org/10.1080/01434632.2014.909445.

DeYoung, C. G., L. C. Quilty, and J. B. Peterson. 2007. "Between Facets and Domains: 10 Aspects of the Big Five." *Journal of Personality and Social Psychology* 93 (5): 880–896. https://doi.org/10.1037/0022-3514.93.5.880.

DeYoung, C. G., Y. J. Weisberg, L. C. Quilty, and J. B. Peterson. 2013. "Unifying the Aspects of the Big Five, the Interpersonal Circumplex, and Trait Affiliation." *Journal of Personality* 81 (5): 465–475. https://doi.org/10.1111/jopy.12020.

Divjak, D. 2019. *Frequency in Language: Memory, Attention and Learning*. New York, NY: Cambridge University Press.

Divjak, D., P. Milin, and S. Medimorec. 2020. "Construal in Language: A Visual-World Approach to the Effects of Linguistic Alternations on Event Perception and Conception." *Cognitive Linguistics* 31 (1): 37–72. https://doi.org/10.1515/cog-2018-0103.

Divjak, D., L. Romain, and P. Milin. 2023. "From Their Point of View: The Article Category as a Hierarchically Structured Referent Tracking System." *Linguistics* 61 (4): 1027–1068. https://doi.org/10.1515/ling-2022-0186.

Donnellan, M. B., F. L. Oswald, B. M. Baird, and R. E. Lucas. 2006. "The Mini-IPIP Scales: Tiny-yet-Effective Measures of the Big Five Factors of Personality." *Psychological Assessment* 18 (2): 192–203. https://doi.org/10.1037/1040-3590.18.2.192.

Dragojevic, M., F. Fasoli, J. Cramer, and T. Rakić. 2021. "Toward a Century of Language Attitudes Research: Looking Back and Moving Forward." *Journal of Language and Social Psychology* 40 (1): 60–79. https://doi.org/10.1177/0261927X20966714.

Dryer, M. S. 1989. "Article-Noun Order." *Proceedings of the 25th Regional Meeting of the Chicago Linguistic Society*, 83–97.

Eisenstein, M. 1983. "Native Reactions to non-Native Speech: A Review of Empirical Research." *Studies in Second Language Acquisition* 5 (2): 160–176. https://doi.org/10.1017/S0272263100004836.

Ensz, K. Y. 1982. "French Attitudes Toward Typical Speech Errors of American Speakers of French." *The Modern Language Journal* 66 (2): 133–139. https://doi.org/10.1111/j.1540-4781.1982.tb06972.x.

Flege, J. E. 1984. "The Detection of French Accent by American Listeners." *The Journal of the Acoustical Society of America* 76 (3): 692–707. https://doi.org/10.1121/1.391256.

Fuertes, J. N., W. H. Gottdiener, H. Martin, T. C. Gilbert, and H. Giles. 2012. "A Meta-Analysis of the Effects of Speakers' Accents on Interpersonal Evaluations." *European Journal of Social Psychology* 42 (1): 120–133. https://doi.org/10.1002/ejsp.862.

Fuertes, J. N., J. C. Potere, and K. Y. Ramirez. 2002. "Effects of Speech Accents on Interpersonal Evaluations: Implications for Counseling Practice and Research." *Cultural Diversity and Ethnic Minority Psychology* 8 (4): 346–356. https://doi.org/10.1037/1099-9809.8.4.347.

Funder, D. C. 2016. "Taking Situations Seriously: The Situation Construal Model and the Riverside Situational Q-Sort." *Current Directions in Psychological Science* 25 (3): 203–208. https://doi.org/10.1177/0963721416635552.

Gaffney, C., and S. Côté. 2019. "Does Personality Influence Ratings of Foreign Accents?" *Journal of Second Language Pronunciation* 6 (1): 68–95. https://doi.org/10.1075/jslp.18042.gaf.

Garrett, P. 2010. *Attitudes to Language*. New York: Cambridge University Press.

Gass, S. M., and A. Mackey. 2017. *Stimulated Recall Methodology in Applied Linguistics and L2 Research*. 2nd ed. New York/London: Routledge.

Giles, H. 1970. "Evaluative Reactions to Accents." *Educational Review* 22 (3): 211–227. https://doi.org/10.1080/0013191700220301.

Giles, H., and P. F. Powesland. 1975. *Speech Style and Social Evaluation*. London: Academic Press.

Gluszek, A., and J. F. Dovidio. 2010. "The way They Speak: A Social Psychological Perspective on the Stigma of Nonnative Accents in Communication." *Personality and Social Psychology Review* 14 (2): 214–237. https://doi.org/10.1177/1088868309359288.

Gnevsheva, K. 2018. "The Expectation Mismatch Effect in Accentedness Perception of Asian and Caucasian non-Native Speakers of English." *Linguistics* 56 (3): 581–598. https://doi.org/10.1515/ling-2018-0006.

Goldberg, L. R. 1999. "A Broad-Bandwidth, Public Domain, Personality Inventory Measuring the Lower-Level Facets of Several Five-Factor Models." In *Personality Psychology in Europe*, edited by I. Mervielde, I. Deary, F. De Fruyt, and F. Ostendorf, Vol. 7, 7–28. Tilburg, The Netherlands: Tilburg University Press.

Grey, S., and J. G. van Hell. 2017. "Foreign-accented Speaker Identity Affects Neural Correlates of Language Comprehension." *Journal of Neurolinguistics* 42:93–108. https://doi.org/10.1016/j.jneuroling.2016.12.001.

Hanulíková, A., P. M. Van Alphen, M. M. Van Goch, and A. Weber. 2012. "When one Person's Mistake is Another's Standard Usage: The Effect of Foreign Accent on Syntactic Processing." *Journal of Cognitive Neuroscience* 24 (4): 878–887. https://doi.org/10.1162/jocn_a_00103.

Hofstee, W. K., B. De Raad, and L. R. Goldberg. 1992. "Integration of the big Five and Circumplex Approaches to Trait Structure." *Journal of Personality and Social Psychology* 63 (1): 146–163. https://doi.org/10.1037/0022-3514.63.1.146.

Hubert Lyall, I., and J. Järvikivi. 2021. "Listener's Personality Traits Predict Changes in Pupil Size During Auditory Language Comprehension." *Scientific Reports* 11 (1): 1–15. https://doi.org/10.1038/s41598-021-84886-3.

John, O. P., and S. Srivastava. 1999. "The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives." In *Handbook of Personality: Theory and Research*, edited by L. A. Pervin, and O. P. John, 2nd ed., 102–138. New York, NY: Guilford Press.

Johnson, N. L. 1949. "Bivariate Distributions Based on Simple Translation Systems." *Biometrika* 36 (3/4): 297–304. https://doi.org/10.2307/2332669.

Johnson, K. 1997. "Speech Perception Without Speaker Normalization: An Exemplar Model." In *Talker Variability in Speech Processing*, edited by K. Johnson, and J. W. Mullennix, 145–165. San Diego, CA: Academic Press.

Johnson, K. 2006. "Resonance in an Exemplar-Based Lexicon: The Emergence of Social Identity and Phonology." *Journal of Phonetics* 34 (4): 485–499. https://doi.org/10.1016/j.wocn.2005.08.004.

Johnson, R., and F. L. Jenks. 1994. Native Speakers' Perceptions of Nonnative Speakers: Related to Phonetic Errors and Spoken Grammatical Errors. Annual Meeting of the Teachers of English to Speakers of Other Languages, March 8-12.

Kang, O. 2012. "Impact of Rater Characteristics and Prosodic Features of Speaker Accentedness on Ratings of International Teaching Assistants' Oral Performance." *Language Assessment Quarterly* 9 (3): 249–269. https://doi.org/10.1080/15434303.2011.642631.

Kang, O., and D. L. Rubin. 2009. "Reverse Linguistic Stereotyping: Measuring the Effect of Listener Expectations on Speech Evaluation." *Journal of Language and Social Psychology* 28 (4): 441–456. https://doi.org/10.1177/0261927X09341950.

Kirjavainen, M., G. Howie, L. A. Rudge, and L. Nicholson. 2024. "The Detriment That Error Production Creates is Affected by non-L1 Speakers' Linguistic Group Membership." *Journal of Second Language Studies* 7 (1): 129–156. https://doi.org/10.1075/jsls.00020.kir.

Klatt, D. H., and L. C. Klatt. 1990. "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers." *The Journal of the Acoustical Society of America* 87 (2): 820–857. https://doi.org/10.1121/1.398894.

Levitt, A., and M. Lucas. 2018. "Effects of Four Voice Qualities and Formant Dispersion on Perception of a Female Voice." *Psychology of Language and Communication* 22 (1): 394–416. https://doi.org/10.2478/plc-2018-0018.

Lin, C., and R. M. Alvarez. 2020. "Personality Traits are Directly Associated with Anti-Black Prejudice in the United States." *PLoS One* 15 (7): e0235436.

Lindemann, S., and N. Subtirelu. 2013. "Reliably Biased: The Role of Listener Expectation in the Perception of Second Language Speech." *Language Learning* 63 (3): 567–594. https://doi.org/10.1111/lang.12014.

Linek, S. B., P. Gerjets, and K. Scheiter. 2010. "The Speaker/Gender Effect: Does the Speaker's Gender Matter When Presenting Auditory Text in Multimedia Messages?" *Instructional Science* 38 (5): 503–521. https://doi.org/10.1007/s11251-009-9115-8.

Magnan, S. S. 1982. "Native Speaker Reaction as a Criterion for Error Correction." In *Report of Central States Conference on the Teaching of Foreign Language*, edited by A. Garfinkel, 30–46. Skokie, IL: National Textbook.

Matthews, G., I. J. Deary, and M. C. Whiteman. 2009. *Personality Traits*. 3rd ed. Cambridge: Cambridge University Press.

Matuschek, H., R. Kliegl, S. Vasishth, H. Baayen, and D. Bates. 2017. "Balancing Type I Error and Power in Linear Mixed Models." *Journal of Memory and Language* 94:305–315. https://doi.org/10.1016/j.jml.2017.01.001.

McGowan, K. B. 2015. "Social Expectation Improves Speech Perception in Noise." *Language and Speech* 58 (4): 502–521. https://doi.org/10.1177/0023830914565191.

Miao, Y., M. Moran, and O. Kang. 2023. "A Cognitive Dissonance Approach to Moderating Listener Perception of L2 English Speakers." *Journal of Multilingual and Multicultural Development*, 1–18. https://doi.org/10.1080/01434632.2023.2236073.

Milin, P., L. B. Feldman, M. Ramscar, P. Hendrix, and R. H. Baayen. 2017. "Discrimination in Lexical Decision." *PLoS One* 12 (2): e0171935. https://doi.org/10.1371/journal.pone.0171935.

Moyer, A. 2021. *The Gifted Language Learner: A Case of Nature Or Nurture?* Cambridge: Cambridge University Press.

Munro, M. J. 2003. "A Primer on Accent Discrimination in the Canadian Context." *TESL Canada Journal* 20 (2): 38–51. https://doi.org/10.18806/tesl.v20i2.947.

Munro, M. J., and T. M. Derwing. 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.

Nejjari, W., M. Gerritsen, M. Van der Haagen, and H. Korzilius. 2012. "Responses to Dutch-Accented English." *World Englishes* 31 (2): 248–267. https://doi.org/10.1111/j.1467-971X.2012.01754.x.

Ortony, A., D. A. Norman, and W. Revelle. 2005. "Affect and Proto-Affect in Effective Functioning." In *Who Needs Emotions? The Brain Meets the Machine*, edited by J.-M. Fellous, and M. A. Arbib, 173–202. New York: Oxford University Press.

Padilla, J. L., and J. P. Leighton. 2017. "Cognitive Interviewing and Think Aloud Methods." In *Understanding and Investigating Response Processes in Validation Research*, edited by B. D. Zumbo, and A. M. Hubley, Vol. 69, 211–228. Cham, Switzerland: Springer International Publishing.

Queen, R., and J. E. Boland. 2015. "I Think Your Going to Like me: Exploring the Role of Errors in Email Messages on Assessments of Potential Housemates." *Linguistics Vanguard* 1 (1): 283–293. https://doi.org/10.1515/lingvan-2015-0011.

R Core Team. 2022. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. https://www.R-project.org/.

Revelle, W. 1997. "Extraversion and Impulsivity: The Lost Dimension?" In *The Scientific Study of Human Nature: Tribute to Hans J. Eysenck at Eighty*, edited by H. Nyborg, 189–212. Oxford: Pergamon/Elsevier Science Inc.

Revelle, W. 2007. "Experimental Approaches to the Study of Personality." In *Handbook of Research Methods in Personality Psychology*, edited by R. W. Robins, R. C. Fraley, and R. F. Krueger, 37–61. New York, NY: The Guilford Press.

Revelle, W. 2021. *How to Use the Psych Package for Mediation/Moderation/Regression Analysis*. http://personality-project.org/r/psych/HowTo/mediation.pdf.

Roessel, J., C. Schoel, and D. Stahlberg. 2020. "Modern Notions of Accent-ism: Findings, Conceptualizations, and Implications for Interventions and Research on Nonnative Accents." *Journal of Language and Social Psychology* 39 (1): 87–111. https://doi.org/10.1177/0261927X19884619.

Romero-Rivas, C., C. Morgan, and T. Collier. 2022. "Accentism on Trial: Categorization/Stereotyping and Implicit Biases Predict Harsher Sentences for Foreign-Accented Defendants." *Journal of Language and Social Psychology* 41 (2): 191–208. https://doi.org/10.1177/0261927X211022785.

Rosseel, Y. 2012. "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software* 48 (2): 1–36. https://doi.org/10.18637/jss.v048.i02.

Rubin, D. L. 1992. "Nonlanguage Factors Affecting Undergraduates' Judgments of Nonnative English-Speaking Teaching Assistants." *Research in Higher Education* 33 (4): 511–531. https://doi.org/10.1007/BF00973770.

Ruivivar, J., and L. Collins. 2018. "The Effects of Foreign Accent on Perceptions of Nonstandard Grammar: A Pilot Study." *Tesol Quarterly* 52 (1): 187–198. https://doi.org/10.1002/tesq.374.

Ruivivar, J., and L. Collins. 2019. "Nonnative Accent and the Perceived Grammaticality of Spoken Grammar Forms." *Journal of Second Language Pronunciation* 5 (2): 269–293. https://doi.org/10.1075/jslp.17039.rui.

Saeli, H. 2018. "Tehrani People's Attitudes Toward Accented Persian: Impact of Gender, age, and Education Level." *International Multilingual Research Journal* 12 (4): 237–254. https://doi.org/10.1080/19313152.2016.1273739.

Soto, C. J., and O. P. John. 2017. "The Next Big Five Inventory (BFI-2): Developing and Assessing a Hierarchical Model with 15 Facets to Enhance Bandwidth, Fidelity, and Predictive Power." *Journal of Personality and Social Psychology* 113 (1): 117–143. https://doi.org/10.1037/pspp0000096.

Spence, J. L., M. J. Hornsey, E. M. Stephenson, and K. Imuta. 2022. "Is Your Accent Right for the Job? A Meta-Analysis on Accent Bias in Hiring Decisions." *Personality and Social Psychology Bulletin* 50 (3): 371–386. https://doi.org/10.1177/01461672221130595.

Tsurutani, C., and E. Selvanathan. 2013. "Influence of Generational Cohort and Experience with non-Native Speakers on Evaluation of Speakers with Foreign-Accented Speech." *International Journal of the Sociology of Language* 2013 (224): 43–62. https://doi.org/10.1515/ijsl-2013-0055

Vann, R. J., D. E. Meyer, and F. O. Lorenz. 1984. "Error Gravity: A Study of Faculty Opinion of ESL Errors." *Tesol Quarterly* 18 (3): 427–440. https://doi.org/10.2307/3586713.

van Rij, J., M. Wieling, R. Baayen, and D. van Rijn. 2017. *itsadug: Interpreting time Series and Autocorrelated Data Using GAMMs (version 2.3)*.

Vasilopoulos, N. L., J. M. Cucina, and A. E. Hunter. 2007. "Personality and Training Proficiency: Issues of Bandwidth-Fidelity and Curvilinearity." *Journal of Occupational and Organizational Psychology* 80 (1): 109–131. https://doi.org/10.1348/096317906X102114.

Wang, Y. A., and M. Rhemtulla. 2021. "Power Analysis for Parameter Estimation in Structural Equation Modeling: A Discussion and Tutorial." *Advances in Methods and Practices in Psychological Science* 4 (1): 1–17. https://doi.org/10.1177/2515245920918253.

Warriner, A. B., V. Kuperman, and M. Brysbaert. 2013. "Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas." *Behavior Research Methods* 45 (4): 1191–1207. https://doi.org/10.3758/s13428-012-0314-x.

Wiggins, J. S. 1996. "An Informal History of the Interpersonal Circumplex Tradition." *Journal of Personality Assessment* 66 (2): 217–233. https://doi.org/10.1207/s15327752jpa6602_2.

Wilt, J., and W. Revelle. 2009. "Extraversion." In *Handbook of Individual Differences in Social Behavior*, edited by M. R. Leary, and R. H. Hoyle, 27–45. New York, NY: The Guilford Press.

Wilt, J., and W. Revelle. 2015. "Affect, Behaviour, Cognition and Desire in the Big Five: An Analysis of Item Content and Structure." *European Journal of Personality* 29 (4): 478–497. https://doi.org/10.1002/per.2002.

Wood, S. N. 2011. "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (1): 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x.

Wood, S. N. 2017. *Generalized Additive Models: An Introduction with R.* Boca Raton, Fla.: CRC press.

Wundt, W., and C. H. Judd. 1897. *Outlines of Psychology.* Oxford, England: Engelmann.

Yerkes, R., and J. Dodson. 1908. "The Relation of Strength of Stimuli to Rapidity of Habit-Information." *Journal of Comparative Neurology and Psychology* 18 (5): 459–482. https://doi.org/10.1002/cne.920180503.

Yoon, K. K. 1993. "Challenging Prototype Descriptions: Perception of Noun Countability and Indefinite vs. Zero Article use." *IRAL-International Review of Applied Linguistics in Language Teaching* 31 (4): 269–290.

Zhu, Y., A. S. L. Fung, and L. Yang. 2021. "A Methodologically Improved Study on Raters' Personality and Rating Severity in Writing Assessment." *SAGE Open* 11 (2): 1–16. https://doi.org/10.1177/21582440211009476.

Appendices

Appendix A. The mini-IPIP scale

No.	Factor	Items
1	E	Am the life of the party.
2	A	Sympathize with others' feelings
3	C	Get chores done right away.
4	N	Have frequent mood swings.
5	I	Have a vivid imagination.
6	E	Don't talk a lot. (R)
7	A	Am not interested in other people's problems. (R)
8	C	Often forget to put things back in their proper place. (R)
9	N	Am relaxed most of the time. (R)
10	I	Am not interested in abstract ideas. (R)
11	E	Talk to a lot of different people at parties.
12	A	Feel others' emotions.
13	C	Like order.
14	N	Get upset easily.
15	I	Have difficulty understanding abstract ideas. (R)
16	E	Keep in the background. (R)
17	A	Am not really interested in others. (R)
18	C	Make a mess of things. (R)
19	N	Seldom feel blue. (R)
20	I	Do not have a good imagination. (R)

Note. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; I = Intellect/Imagination; (R) = Reverse Scored Item.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. *Psychological assessment, 18*(2), 192-203.

Appendix B. Power analysis

Based on the research design, a prototype model was specified, containing dependent and predictor variables used in the analysis. The prototype model comprised all main and interaction effects, linear and quadratic. By including all potential effect in the prototype model, we aimed to create sub-optimal conditions for model estimation: specifying the effects that were not likely to simultaneously be estimated as large and statistically significant. This allowed us to obtain the most conservative power analysis. The model analysis consisted of several, with systematic variation in sample sizes and effect sizes.

Values for small, moderate and large effects were chosen according to the Monte Carlo simulation by Wang and Rhemtulla (2021). In Stage one, all population effects have been set to zero. In Stage two, the population effects were small (0.1), in Stage three moderate (0.3), and in Stage four large (0.5). According to these specifications, simulated populations ($N = 24$ million) were generated randomly to reflect model features with the respective effect sizes. The rest of the procedure was the same for all stages. From the population of $N = 24M$, 1000 randomly drawn samples were extracted of the following sizes: $n_1 = 30$, $n_2 = 60$, $n_3 = 90$, $n_4 = 120$, $n_5 = 180$, $n_6 = 240$; samples were with replacement. For each of the 6000 random sample (i.e. 1000 for each set sample size), the prototype model was fit using the lavaan statistical package (Rosseel 2012) for the R software environment for statistical computing (R Core Team 2022) and effects and their *p*-values were extracted. Given that the stimuli were repeated across

participants, participant ID was specified as the 'cluster' variable, or the random effect, in the analyses. See the Supporting Information for the results of different sample sizes and effect sizes alongside their p-values.

With zero effects in the population, in samples $n_1 = 30$, approximately 92% of parameters were estimated as ranging between 0 and 0.1 and being statistically insignificant. The proportions were virtually identical (though raw values differed slightly) for all sample sizes.

For small (0.1) population effects, in samples $n_1 = 30$ around 34% were estimated to be between 0.1 and 0.3, with p-values below .01. Additional 37% estimates ranged from 0 to 0.1, with p-values below .05. In samples $n_2 = 60$, 78% parameters were estimated as ranging from 0 and 0.3, with p-values up to .05. In samples $n_3 = 90$, about 91% effect values spanned from 0 to 0.3, with p-values at .05 or lower.

When population effects were set to medium, in samples $n_1 = 30$, 75% estimates spanned from 0.1–0.5 with p-values below .001, with additional 16% ranging from 0 to 0.1 ($p \leq .001$) and 5% estimated at above 0.5 ($p \leq .001$). Similar proportions were obtained for the samples sized $n_2 = 60$ and $n_3 = 90$.

For large population effects, in samples $n_1 = 30$, 96% were estimated as larger than 0.1, at p-values of .001 or smaller. Similar proportions occurred in the samples sized $n_2 = 60$ and $n_3 = 90$.

The results suggest that, in sub-optimal conditions, the risk of type I error is rather small, i.e. non-existing population parameters were estimated as statistically non-significant. As for the estimation of the existing population effects, approximately 80% small effects were estimated correctly for samples at $n_2 = 60$ or less. For medium effects, the proportion of correct estimates already exceeds 90% in samples $n_1 = 30$. Thus, we may conclude that our sample size ($N = 60$) is sufficient to ensure acceptable statistical power in detecting small effects and satisfactory or excellent statistical power in detecting medium to large effects.

Appendix C. Data preparation

Data exploration revealed that the dependent variables were characterised by three properties that make model fitting challenging: (1) a substantial number of acceptability scores were at ceiling, i.e. a number of speech stimuli received the maximum score of 100; (2) the distribution was skewed; and (3) the variability between the four accent-grammaticality conditions differed. We will discuss these issues in turn.

The maximum scorers included ratings of both British (203 datapoints for well-formed, 63 datapoints for error-filled) and Polish (110 datapoints for well-formed, 29 datapoints for error-filled) accented stimuli, making up about 22% and 12% of the total data respectively (1200 datapoints for each accent condition). This observation is interesting in and of itself. Although half of the British accented stimuli did not contain any errors and hence should have been rated as perfectly acceptable (600 datapoints), less than half of them received maximum acceptability scores. Instead, 18% of well-formed Polish accented stimuli were perceived as fully acceptable, so were a small number of error-filled stimuli.

On the other hand, this proportion of maximum scores causes skewness in the distribution of the data and is potentially detrimental to the variability of the dependent variable. As shown in Figure A1, the acceptability judgment scores revealed a strongly left-skewed distribution and additionally some bimodality (see the cutting lines), instead of a normal distribution. Fligner-Killeen tests (a robust non-parametric test for the homogeneity of k variances; cf., Conover, Johnson, and Johnson 1981) on the acceptability scores, taking into account the speaker accent-grammaticality conditions (British vs. Polish, error-filled vs. well-formed), turned out to be highly significant and, thus, confirmed the non-homogeneity of variances across the four conditions ($\chi^2 = 174.27$, $df = 3$, $p < 0.0001$). As Figure A2 illustrates, the variability in the acceptability scores for each of the four conditions differs: B0 is consistently judged very highly, while B1 and in particular P1 display high variability. Such differences in variability between groups create problems for statistical analysis.

A reasonable and cautious approach to such data could consist of two types of analyses: a logistic model to predict differences between max vs. sub-max acceptability scores, and a (version of)linear model for a closer look at the results of the sub-max scores, as they appear to generate a more normal-like distribution. Unfortunately, however,the binary split (a dichotomy) between the max and sub-max scores was rather uneven (roughly a 20:80 split), which would make a logistic analysis unreliable.Therefore, we proceeded with the analysis of the ratings of sub-max scores, i.e. those that did not score the maximum of 100.
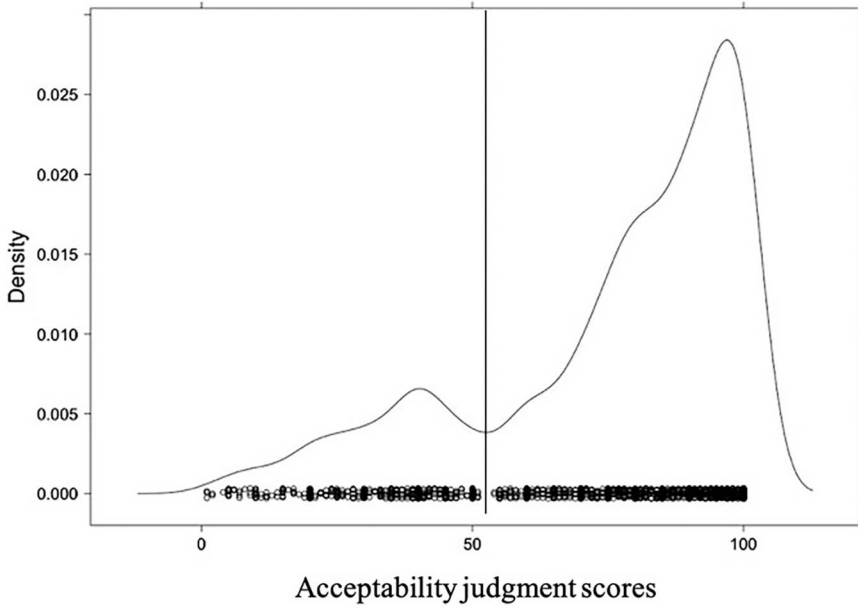
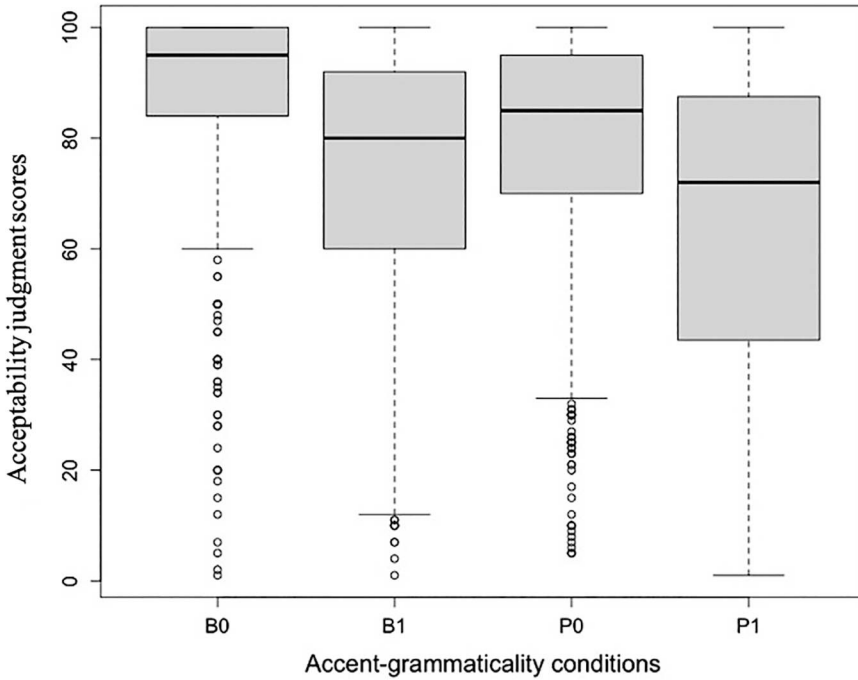**Figure A1.** Density plots of acceptability judgment scores.

Figure A2. Boxplot of acceptability judgment scores on speech stimuli acrossfour conditions. B.0 = British accent, without errors; B1 = British accent,with errors; P0 = Polish accent, without errors; P1 = Polish accent,with errors.