To be published in *Annual Review of Applied Linguistics* (Cambridge University Press)

Developing, Analyzing and Sharing Multivariate and Multifactorial Datasets for Open Science: Individual Differences in the Dynamic System of L2 Speech Learning Revisited

Kazuya Saito (*University College London*)[1]

Konstantinos Macmillan (*Birkbeck, University of London*)

Tran Mai (*Birkbeck, University of London*)

Yui Suzukida (*University College London*)

Hui Sun (*University of Birmingham*)

Viktoria Magne (*University of West London*)

Meltem Ilkan (*Birkbeck, University of London*)

Akira Murakami (*University of Birmingham*)

Abstract

Following the trends established in psychology and emerging in L2 research, we explain our support for an Open Science approach in this paper (i.e., developing, analyzing and sharing datasets) as a way to answer controversial and complex questions in applied linguistics. We illustrate this with a focus on a frequently debated question, what underlies individual differences in the dynamic system of post-pubertal L2 speech learning? We provide a detailed description of our dataset which consists of spontaneous speech samples, elicited from 110 late L2 speakers in the UK with diverse linguistic, experiential and sociopsychological backgrounds, rated by ten L1 English listeners for comprehensibility and nativelikeness. We explain how we examined the source of individual differences by linking different levels of L2 speech performance to a range of learner-extrinsic and intrinsic variables related to first language backgrounds, age, experience, motivation, awareness, and attitudes using a series of factor and Bayesian mixed-effects ordinal regression analyses. We conclude with a range of suggestions for the applied linguistics and SLA, including the use of Bayesian methods in analyzing multivariate, multifactorial data of this kind, and advocating publicly available datasets. In keeping with recommendations for increasing openness of the field, we invite readers to rethink and redo our analyses and interpretations from multiple angles by making our dataset publicly available and coding as part of our 40th anniversary ARAL article.

Key words: Open science, individual differences, L2 speech, comprehensibility, Bayesian methods

[1] This study was funded by Leverhulme Trust Research Grant (RPG-2019-039) and Arnold Bentley New Initiatives Fund. We would like to thank the anonymous reviewers from the *Annual Review of Applied Linguistics* and Editor Alison Mackey for their useful and constructive comments on the earlier versions of the manuscript. Corresponding Author: Kazuya Saito (k.saito@ucl.ac.uk).

Introduction

In the field of applied linguistics and second language acquisition (SLA), a growing number of scholars have emphasized the importance of the Open Science approach (e.g., Marsden, in press). One crucial component of this movement is to make all the research processes related to data collection and analysis fully transparent. As such, readers can not only understand exactly what the researchers attempted to do, but also conduct objective and independent replications of the findings in the future. Such an approach is particularly important when it comes to theoretically and practically crucial topics that need to be replicated in many different contexts. In this paper, we aim to demonstrate how the Open Science approach allows us to consider a fundamental, yet controversial issue—i.e., why the rate and ultimate attainment of L2 learners is so varied, especially when they start learning a target language after puberty.

Over the past 50 years, the role of individual differences in post-pubertal L2 speech learning has attracted a great amount of scholarly attention. While many demonstrate detectable L1-related accents even after years of practice, some L2 learners can attain highly advanced L2 pronunciation proficiency (e.g., Flege, Munro, & MacKay, 1995). To examine the source of such variation, this line of L2 speech research has traditionally considered only one or two individual difference variables (e.g., age, motivation) at a time. More recently, scholars have begun to describe L2 learning as a complex, adaptive, emergent, self-organizing, and ever-changing system (e.g., Larsen-Freeman, 2012). To unravel what underlies a dynamic phenomenon of this kind, we argue that it is crucial to include as many learner-internal and learner-external factors as possible within the same research design. In addition, prior work has typically assessed L2 speech proficiency in terms of the degree of nativelikeness (or accentedness). However, the levels of attainment in post-pubertal L2 pronunciation should be assessed based on ease of understanding (comprehensibility), because many adult L2 learners can be highly comprehensible despite their detectable L2 accents (Derwing & Munro, 2013; Saito, Trofimovich, & Isaacs, 2017).

Considering all the methodological concerns above (i.e., the lack of data transparency, depth and diversity), the primary objective of the current study is to revisit the process and product of late L2 speech learning. Our study is novel, as we consider the notion of the dynamic system (i.e., simultaneous consideration of multiple dependent and independent variables) and the Open Science approach (i.e., developing, analyzing, and sharing dataset

with interested audience). First, we report in detail how we constructed a relatively large-scale learner and speech dataset among 110 late L2 speakers in London. Subsequently, we present the results of regression modeling analyses to shed light on what types of learner variables, related to the learners' L1, age, experience, motivation, awareness and attitudes, *jointly* interact to determine different levels of L2 nativelikeness *and* comprehensibility. Last, we make the actual dataset publicly available while providing a range of suggestions regarding how to analyze multivariate, multifactorial data of this kind, and inviting the readers to rethink our analyses and interpretations from multiple perspectives (see **DATASET**).

Background

Many early bilinguals attain high levels of L2 proficiency through mere exposure to the target language in an implicit and incidental fashion (like in L1 acquisition). With respect to late L2 speakers, who start learning a target language after puberty, their speech is generally L2-accented as it builds on and interacts with their already-developed L1 system (Flege et al., 1995). The degree of such foreign accentedness can vary greatly due to a range of learner-external (L1-L2 distance, age, experience) and -internal factors (motivation, awareness, attitudes). To date, previous studies have typically looked at one or two independent variables in isolation, and linked them to the nativelikeness of participants' L2 speech performance.

External Factors of L2 Speech Learning

L1-L2 Distance. A range of theoretical accounts have been proposed to explain the influence of L1 phonetic structures on L2 speech learning. A core premise of such accounts is that the linguistic distance between an L1 and L2 determines pronunciation learning difficulty (Best & Tyler, 2007, for Perceptual Assimilation Model). Numerous empirical studies have documented learners' difficulty in acquiring relatively new articulatory and acoustic features in an L2 on segmental (e.g., Japanese learners' English /r/-/l/ acquisition; Saito, 2013) and suprasegmental (e.g., American learners' Mandarin lexical tone acquisition; Wang, Jongman, & Sereno, 2003) levels. Conversely, there is some evidence that even late L2 learners can attain highly advanced L2 pronunciation proficiency especially when their L2 is linguistically

close to their L1 (e.g., Bongaerts, van Summeren, Planken, & Schils, 1997, for Dutch learners of English).

Age. To date, scholars have extensively examined the extent to which L1 influence could be mediated by a set of age-related factors, such as the age of arrival (i.e., the first exposure to the target language in a naturalistic setting), age of learning (i.e., the onset of foreign language education) and testing (i.e., participants' age at the time of data collection). Although age of acquisition has been found to predict the ultimate attainment of L2 oral proficiency after years of immersion in an L2-speaking environment (e.g., Flege et al., 1995; Saito, 2013), the predictive power of age has remained ambiguous in the context of foreign language learning (several hours of form-focused instruction per week). The existing literature has pointed out that late starters may benefit more from foreign language instruction due to their cognitive maturity, fully developed L1 literacy and accumulative classroom experience (e.g., Muñoz, 2014).

Experience. Another variable relevant for late L2 speech learning is concerned with quantity (how much learners have practiced) and quality (how, with whom and what learners have practiced) of experience. Length of residence (LOR) in an L2 environment has been adopted in L2 speech research as a proxy for the quantity of L2 use; however, the reliability of LOR has been subject to criticism because the frequency of L1 and L2 use differs greatly among individuals, even if they stay in an L2 speaking environment for the same period of time (for more relevant discussion, see Saito, 2015). In this regard, scholars have looked at the quality of experience from multiple angles, such as the ratio of language use (L1 vs. L2) (e.g., Flege, MacKay, & Piske, 2002), type of interlocutors (fluent vs. non-fluent speakers) (e.g., Muñoz & Llanes, 2014), and context of interaction (social vs. professional vs. family) (e.g., Jia & Aaronson, 2003).

Learner-Internal Factors of L2 Speech Learning

Metalinguistic awareness. From a theoretical standpoint, awareness (i.e., explicit knowledge about target language) is believed to play a key role in L2 acquisition, because it helps L2 learners to better notice and understand specific features in received input, and then internalize them into long-term memory (Schmidt, 2001). A series of experimental studies

have convincingly shown that L2 learners exhibit some gains when they practice an L2 explicitly, consciously and deliberately (e.g., Hama & Leow, 2010). In terms of L2 phonology, there is some evidence that L2 learners with greater phonological awareness (i.e., conscious knowledge about phonological and phonetic structures of a target language) tend to produce not only more segmentally accurate (Saito, 2019), but also more comprehensible speech (Venkatagiri & Levis, 2007).

Motivation. In other studies, highly advanced L2 speakers have been reported to demonstrate high levels of professional and integrative *motivation* to use language accurately, under various circumstances (school, business, social, and home-related). For example, such speakers may be L2 language teachers by profession (Bongaerts et al., 1997) and/or have intensive immersion experience through international marriages (Ioup, Boustagi, El Tigi, & Moselle, 1994) (for the role of future visions and relevant motivation, see Saito, Dewaele, Abe, & In'nami, 2018).

Attitude. Another well-researched topic is concerned with attitudes, defined as "an evaluative orientation to a social object" (Garrett, 2010, p. 3). Whereas scholars have extensively examined language attitudes toward L2 learning and teaching in general (see Gardner & Smythe, 1981, for the influential framework and Attitude/Motivation Test Battery), some studies have looked at this topic in the context of L2 pronunciation. For example, previous research has shown that some L2 learners express solidarity with their L1-accented speech, which translates into positive attitudes toward speakers from the same L1 background (McKenzie & Gilmore, 2017). In the context of French-speaking Quebec, Gatbonton and Trofimovich (2008) found that strong L1 ethnic group affiliation was associated with low L2 proficiency, whereas positive views toward both L1 and L2 communities were linked to high L2 pronunciation proficiency.

Comprehensibility vs. Nativelikeness

Importantly, much of the late L2 speech literature has been exclusively concerned with the relationship between learners' extrinsic and intrinsic individual differences, and the degree of L2 phonological *nativelikeness*. In the field of SLA, however, there has been a consensus that the linguistic behaviors of bilinguals and monolinguals are essentially different; and that L2 speakers' linguistic performance should be compared within themselves instead of in comparison with an idealized monolingual native speaker model (e.g., Ortega, 2018). In line with this paradigm shift, a growing number of scholars have emphasized the importance of examining L2 speech from the perspective of comprehensibility rather than nativelikeness (Derwing & Munro, 2013; Saito et al., 2017).

To date, many empirical studies have indeed shown that perceived comprehensibility and nativelikeness tap into somewhat overlapping but essentially different constructs of L2 speech. For example, while assessing the comprehensibility of L2 speech, listeners are found to attune to a range of linguistic elements, especially those directly relevant to successful comprehension, in order to arrive at the overall meaning of L2-accented speech in the most efficient and effective fashion (e.g., Suzuki & Kormos, 2019, for prosody). L2 learners can continue to enhance the comprehensibility of their speech regardless of detectable L2 accents, as long as they regularly use their L2 for social interaction with various fluent speakers in diverse social settings (Derwing & Munro, 2013). In contrast, listeners tend to assess the degree of linguistic nativelikeness solely based on phonological accuracy (Saito, Trofimovich, & Isaacs, 2016); the perceived nativelike aspects of L2 speech is resistant to change, especially after the initial rapid development within the first few years of immersion (Saito & Munro, 2014).

Open Science Approach

With the aim of attaining scholarly rigor, the importance of Open Science has been extensively discussed in various academic disciplines (for an overview, see McKiernan et al., 2016). It has been increasingly adopted as a mandatory condition for authors publishing work in major academic journals (e.g., Gewin, 2016, for *Nature*; Gerrig & Rastle, 2019, for *Journal of Memory and Language*; Marsden, Crossley, Ellis, Kormos, Morgan-Short, & Thierry, 2019, for *Language Learning*). The Open Science approach refers to a wide range of

research practices, which include depositing academic literature in freely available platforms (open repository), creating an accessible summary for the general public (open access), and sharing all research materials and datasets (open data). Importantly, the benefits of such open practices are compelling: such as boosting citations, media attention, potential collaborators, and funding opportunities (see McKiernan et al., 2016).

Despite its popularity in diverse areas of science, the Open Science approach to research has been significantly lacking in the field of SLA (Marsden, in press). While the number of meta-analyses has been increasing, many primary studies were reported to be eliminated due to the unavailability of data, indicating that the findings of these studies may not necessarily reflect the state-of-the-art status of the field (Larson-Hall & Plonsky, 2015). Relatedly, recent methodological synthesis papers have revealed that a very small portion of individual studies made their materials available (e.g., Marsden, Thompson & Plonsky, 2018, for 4% out of 71 self-paced reading studies; Plonsky, Marsden, Crowther, Gass & Spinner, 2019, for 35% out of 214 grammatical judgement studies). These problems subsequently hinder third party researchers from examining the replicability and generalizability of existing research findings (Marsden, Morgan-Short, Thompson, & Abugaber, 2018).

Motivation for Current Study

Whereas a growing number of scholars have accepted the view that L2 speech is a multifaceted phenomenon, existing research has been mainly concerned with how one or two independent variables could affect the outcomes of L2 speech. Unfortunately, this line of work fails to see L2 learning as a complex dynamic system (Larsen-Freeman, 2012). We have yet to determine how a range of different learner-external and learner-internal factors *jointly* interact to influence the rate and ultimate attainment of late learners' L2 pronunciation. Such research will shed light on our understanding of what accounts for linguistic, experiential and sociopsychological underpinnings of late L2 speech learning, as well as informing future practices how to best help different types of learners who aim to achieve comprehensible L2 pronunciation vs. those who strive to achieve nativelike L2 pronunciation. Our research question, therefore, is as follows:

- How do learner-external and learner-internal factors differentially relate to L2 learners' speech comprehensibility and nativelikeness?

In order to answer this research question, we took two unique approaches—including numerous independent and dependent variables to examine L2 speech as a dynamic system (i.e., the dynamic perspective), and constructing, analyzing and sharing the entire dataset (i.e., the Open Science approach).

In the context of 110 late L2 learners in London, we first explicate what kinds of profiles characterize L2 learners who have achieved varying levels of L2 comprehensibility and nativelikeness. Following the notion of the Open Science approach, therefore, we provide all the details in terms of what research instruments we used to collect the dataset (speaking test, learner questionnaire, rater training scripts), what kinds of statistical analyses we adopted (data reduction, mixed-effects modeling), and how we interpreted the findings. In order to test the scientific rigor of the current study, we would like to invite the readers not only to replicate the method that we developed, and reproduce the results that we reached, but also to *critically* look at the way we operationalized the current project and think of different types of statistical analyses to approach the dataset with, i.e., the strong version of data transparency (Marsden, in press).

Method

Dataset

Given that the scope of the study highlights late L2 learners, we carefully focused on late L2 learners whose age of arrival at an English-speaking environment was beyond the age of 16. These learners were assumed to speak L2 English with perceptible L1-related accents (for a similar definition, see Flege et al., 1995). To recruit a sufficient number of L2 speakers that could represent a wide range of L2 oral proficiency levels (beginner to advanced), flyers were circulated at various locations (universities, language schools) and on social media. All data collection took place individually in a quiet room at the participants' residence, offices, schools, and community centers for their convenience. For each session, participants were first interviewed to gather a range of information related to their L1 backgrounds, age, experience, motivation, awareness and attitudes (see **Supporting Information-A** for the full-length questionnaire). This was followed by a speech recording session, wherein the participants' spontaneous speech was elicited via a timed picture description task.

The participants widely differed vis-à-vis a total of 30 learner variables spanning L1 backgrounds, age of acquisition, language quantity and quality of experience, professional and social motivation, and awareness and attitudes toward foreign-accented vs. nativelike speech. For the raw data and descriptive statistics of the 30 variables, see **DATASET** and **Supporting Information-B**.

- **First Language Backgrounds (1 variable):** The participants in the current study were classified into nine major language families: (1) Romance (*n* = 19) (e.g., Italian, Spanish, French), (2) Germanic (*n* = 5) (e.g., German, Swedish, Dutch), (3) Indo-Iranian (*n* = 4) (e.g., Hindi-Urdu, Bengali, Punjabi), (4) Balto-Slavic (*n* = 18) (e.g., Russian, Polish, Czech), (5) Uralic languages (*n* = 2) (Estonian), (6) Sino-Tibetan (*n* = 15) (Chinese), (7) Altaic (*n* = 25) (Japanese, Korean, Turkish), (8) Austro-Asiatic (*n* = 12) (Vietnamese), and (9) Niger-Congo (*n* = 10) (Yoruba, Igbo, Swahili). For the analyses, the dummy code—"0" Indo-European (*n* = 46); "1" non-Indo-European (*n* = 64)—was used to see how the L1-L2 distance could be associated with the comprehensibility of their L2 English speech.

- **Age (3 variables):** The participants' age profiles were substantially different in terms of age of arrival at an English-speaking environment (i.e., age of acquisition) (*Range* = 16–55), the onset of foreign language education (i.e., age of learning) (*Range* = 2–58) and data collection (i.e., age of testing) (*Range* = 20–59).

- **Previous Experience (5 variables):** In the current study, participants' previous experience was surveyed in terms of (i) how long they had practiced English in foreign language classrooms (*Range* = 0–23 years) and (ii) how long they had stayed in English speaking countries (*Range* = 0.1–39 years). Approximately 30% of the participants reported previous experience in (iii) linguistics training (*n* = 33) or/and (iv) teaching English as an L2 (*n* = 31). We also created (v) a composite, broad category to capture the number of participants who had received any type of professional training related to linguistics or/and teaching (*n* = 36).

- **Current Experience (9 variables):** To scrutinize *current* experience in the UK, following the questionnaire format of the Language Contact Profile (Freed, Dewey,

Segalowitz, & Halter, 2004), participants were asked to self-report the percentage of time they spent using their L1 and L2 (English in this case) at the time of the project. As per three different settings: professional (work/school), social (with friends) and home (with family). To further examine the type of interlocutors, the participants were asked to estimate the percentage of time they spent interacting in L2 English with fluent vs. non-fluent speakers.

- **Motivation (3 variables):** There is some evidence that very few L2 learners attain near-nativelike pronunciation. Such learners often demonstrate strong concern for the attainment and use of high-level L2 proficiency due to their profession (Bongaerts et al., 1997; Flege et al., 1995) and communication with family members (Ioup et al., 1994). The participants rated the degree to which they were expected to use L2 English at a nativelike proficiency level on a 9-point scale (*1 = not at all, 9 = very much so*) for three different settings: professional (work/school), social (with friends), and home (with family).[2]

- **Awareness (5 variables):** Following the methodological practices in L2 awareness research (e.g., Hama & Leow, 2010), the participants' awareness of L2 comprehensibility was measured via self-reports. In the current study, we interviewed the participants to find out the extent to which they were aware of the importance of specific linguistic dimensions in L2 speech. Participants rated which aspects of language they thought were relatively crucial for successful communication on a 9-point scale (*1 = not important, 9 = very important*). The five statements included were: (a) speaking English without any accent like a native speaker; (b) speaking comprehensible English regardless of accentedness; (c) good pronunciation; (d) appropriate vocabulary and grammar, and (e) idiomatic and sophisticated expression.

- **Familiarity and Attitudes (4 variables):** In the current study, the participants' familiarity and attitudes (i.e., perception) toward foreign-accented and nativelike pronunciation were measured via their self-ratings of the four statements on a 9-point

[2] In previous L2 pronunciation studies, the same methodology (self-ratings) has been used to measure learners' awareness of various aspects of L2 pronunciation but using different labels for the phenomenon that researchers were examining (e.g., Elliott, 1995 for "concern for pronunciation accuracy").

scale (*1 = strongly disagree, 9 = strongly agree*). For familiarity, the two statements asked the extent to which participants were familiar with different types of L2 accented English and British English. For attitudes, the other statements asked the extent to which participants liked it when people speak English with a foreign accent; and with a British accent (for a similar method, see Gatbonton & Trofimovich, 2008).

Comprehensibility and Nativelikeness Judgements

Speaking Materials. In previous L2 speech studies, word, sentence and paragraph reading tasks have often been adopted as outcome measures. However, the construct validity of such controlled tasks has remained controversial, because its format allows adult L2 learners to carefully monitor their correct pronunciation forms without much communicative pressure. In order to index adult L2 learners' pronunciation proficiency, a growing number of scholars have emphasized the importance of adopting *spontaneous* speech tasks, wherein speakers' primary focus lies in conveying the intended message while simultaneously paying attention to phonological, lexical, grammatical and discoursal aspects of language (Piske, Flege, MacKay, & Meador, 2011; Saito & Plonsky, 2019).

Based on this rationale, a decision was made to use a timed picture description task to elicit certain lengths of spontaneous speech without too many disfluencies from L2 learners with *varied* proficiency levels (beginner to advanced). The participants described seven different pictures under time pressure (five seconds of planning per picture). To avoid false starts and to support true beginners, participants were instructed to use three given key words relevant to the content of each picture (for a similar spontaneous task modality, see Munro, 2013 for a picture-naming task). To control for task familiarity, the first four picture descriptions were used as practice, and the last three descriptions were submitted to final analyses. There was no time limit for each picture description. All speech samples were individually recorded via a portable MP3 recorder, and normalized for peak amplitude. The first 10 seconds of the three picture descriptions were cut and stored as one single MP3 file per participant, with each participant contributing roughly 30 seconds of spontaneous speech. The length of speech per participant could be considered sufficiently long to provide raters with enough linguistic information in conjunction with the standard in L2 speech research (Hopp & Schmid, 2013, for 10-20 seconds; Derwing & Munro, 2013 for 30 seconds). The task instruction and materials were deposited in IRIS (Marsden, Mackey, & Plonsky, 2016).

Raters. A total of ten native speaking raters (6 males, 4 females) were recruited in London (M_{age} = 19.5 years). All of them reported that at least one of their parents/carers was an L1 English speaker; and that they used English as their primary language of communication in professional, social and home settings ($M_{\%\ of\ English\ use\ per\ day}$ = 99.0%). Since the raters were living in London (a highly multilingual city) at the time of the project, they reported relatively high levels of familiarity with foreign-accented speech (M = 5.2; *1 = not at all, 6 = very much*). None of them reported having prior linguistics training nor hearing problems.

Procedure. All the rating sessions took place individually in a quiet room at a university in London. The speech samples were played in a randomized order via PRAAT (Boersma & Weenink, 2017). Upon hearing each sample, raters were asked to assess them on a 9-point scale for comprehensibility (*1 = very difficult to understand, 9 = very easy to understand*) and nativelikeness (*1 = not native-like, 9 = completely native-like*). Since L2 comprehensibility and nativelikeness, by definition, involves "intuitive" judgements, raters were only able to listen to each sample once (no replay button was available).

Raters first received a brief explanation of comprehensibility and nativelikeness from a trained researcher, and how to make their ratings (see **Supporting Information-C** for training scripts). After familiarizing themselves with the picture prompts used to elicit speech, they practiced the rating procedure by using three representative samples which were not included in the main dataset (beginner, intermediate, advanced). Then, the raters proceeded to the main dataset (N = 110 L2 speakers). Raters took a five-minute intermission halfway through. An entire session lasted for approximately two hours. For the raters' comprehensibility and nativelikeness scores, see **DATASET**.

Statistical Analysis Procedure

There were two potential issues in the examination of the relationship between the characteristics of the learners identified in the above manner, and their L2 English speech comprehensibility and nativelikeness. First, the number of learner variables (n = 30) was fairly large, considering the number of learners (N = 110). Since our goal was to explain between-learner variability, the former should be much smaller than the latter. Secondly,

some of the 30 learner individual difference variables were highly correlated, which could, in turn, make it difficult to separate their effects. To further reduce the number of predictor variables, all the learner variables were submitted to a factor analysis to identify latent variables underlying the 30 elicited learner variables. The factor scores were then submitted to a regression model to investigate the relationship between the factors and L2 speech comprehensibility and nativelikeness scores.

Results

Underlying Learner Variables

The first objective of the statistical analyses was to examine a number of underlying factors among a total of 30 learner variables related to the participants' L1 background, age, experience, motivation, awareness and attitudes. Following Loewen and Gonulal's (2015) field-specific guidelines for analyzing factorability and determining a threshold for factor loadings, participants' questionnaire data was submitted to a factor analysis with Direct Oblimin rotation and the principal component extraction method. Loewen and Gonulal pointed out that the cumulative percentage of explained variance reported in L2 research is relatively low (60-65%). To increase the cumulative percentage of explained variance (> 80%), the Jolliffe criterion was adopted with the eigenvalue set to 0.8. Two tests were conducted to confirm the factorability of the entire dataset: the Bartlett's test of sphericity and the Kaiser-Meyer-Olkin measure of sample adequacy. To select the practically significant factor loadings, 0.5 was used as the cut-off value.

The first model identified 13 factors capturing 82.3% of the variance among the 30 learner variables. Although the Bartlett's test was significant ($\chi^2 = 2067.542$, $p < .001$), the Kaiser-Meyer-Olkin (KMO) value was relatively low (i.e., .419), suggesting that the sampling of the dataset is questionable. According to our inspection of the pattern matrix, one obvious confusion was related to the nine current experience variables which showed a set of strong correlations with each other ($r = .3$-$.8$). Some variables were not clearly clustered into any overall factors (e.g., L1 use at work). To enhance the factorability of the dataset, we reduced the nine experience variables into two averaged scores per participant by averaging across the following subcategories across all different contexts (work, social, home): (a) how

much they were using their L1 and (b) how much they were using their L2 with fluent users (including L1 speakers and advanced L2 speakers).

The second model identified 11 factors explaining 82.5% of the variance among the 23 learner variables. We considered the factorability to be adequate according to the results of the Bartlett's test ($\chi^2 = 1226.456$, $p < .001$) and KMO test (.547). In conjunction with the pattern matrix summarized in **Supporting Information-D**, each factor was labelled as follows:

- Factor 1 was labeled as "**Experience Quantity**" as the items with high loadings concerned the extent to which participants had been in L2 English speaking environments prior to the project.
- Factor 2 was labelled as "**Current L2 Use**" as it covered two variables related to the extent to which L2 learners used L2 (instead of L1) especially with fluent speakers at the time of the project.
- Factor 3 was labeled as "**Awareness of Nativeness**" as the items clustered here indexed the extent to which participants perceived the importance of nativelike use of language, phonology and idiomatic expressions.
- Factor 4 was labeled as "**Age of Immersion**" as it clustered all the timing variables such as the age of arrival in English speaking countries.
- Factor 5 was labeled as "**Motivation**" as it featured all the items related to participants' motivation and concern for nativelike English pronunciation in different settings.
- Factor 6 was labeled as "**Attitude to Nativeness**" as it reflected the extent to which they appreciated, preferred, and had been familiarized with British English.
- Factor 7 was labeled as "**EFL Experience**" as it featured how early they had started learning English in the classroom setting, and for how long they had received foreign language education prior to their arrival in English countries.
- Factor 8 was labeled as "**Special Past Experience**" as it spotted participants who had previously received linguistics training and/or L2 English teaching experience.
- Factor 9 was labeled as "**Attitude to Foreign Accents**" as it captured only one learner variable (i.e., the extent to which participants liked it when others spoke English with a foreign accent).

- Factor 10 was labeled as "**Comprehensibility Orientation**" it covered not only how much participants were familiar with foreign-accented English, but also the extent to which they perceived the importance of comprehensibility in successful L2 communication.
- Factor 11 was labeled as "**L1 Influence**" as it corresponded to the extent to which participants' L1 background is far from/close to L2 English (i.e., Indo-European language).

Factor scores were then computed with the Bartlett's method and their relationships with comprehensibility/nativelikeness ratings were visualized and analyzed (see **Supporting Information-E)**.

Regression Modeling

In order to formally investigate the relationship between the factor scores of the 11 factors identified, and the ratings of L2 comprehensibility and nativelikeness, we employed a Bayesian multivariate mixed-effects ordinal regression model. We opted for a Bayesian approach because it (a) allows us to estimate the full posterior distribution, which is more informative than the frequentist point estimate (Kruschke, 2014), (b) generates more intuitive metrics of uncertainty (Lambert, 2018), and (c) employs the tools that allow flexible and complex modeling (e.g., Carpenter et al., 2017). Readers are referred to Lambert (2018) and Kruschke (2014) for an accessible introduction to Bayesian data analysis, as well as to Norouzian, de Miranda, and Plonsky (2018) for field-specific recommendations on the use of the Bayesian approach.

Multivariate models permit the simultaneous modeling of multiple outcome variables, such as the two kinds of ratings in the present study (see Hui, 2019). Furthermore, comprehensibility and nativelikeness ratings consist of ordered categories, and analyzing an ordinal variable with techniques assuming continuous variables causes several problems (Liddell & Kruschke, 2018). Therefore, in the present study an ordinal regression was employed. The statistical models were fit with brms (Bürkner, 2017), a front-end R package of Stan (Carpenter et al., 2017). The R code is available (see **RCODE**).

Among multiple classes of ordinal models, we employed a cumulative model, which assumes continuous variables underlying our observed rating variables (Bürkner & Vuorre,

2019). The error term was assumed to follow a logistic distribution. The model specifically included individual ratings of comprehensibility and nativelikeness as dependent variables, 11 sets of factor scores as fixed-effects variables, and by-learner and by-rater random intercepts. The correlation between the random intercepts of comprehensibility and those of nativelikeness ratings was modeled within each random-effects factor (i.e., learner and rater). No interaction term or random slope was included due to a relatively large number of predictors for the given number of learners. Nonlinear effects were not examined for the same reason. Variable selection was not performed due to the many issues associated with the procedure (Harrell, 2015).

For all the parameters, weakly informative prior distributions were used. Specifically, (i) standard normal distributions were specified for slope coefficients representing the effects of each factor, (ii) student-t distribution with the mean of zero, the degree of freedom of three, and the scale of 10 were used for the parameters representing the threshold values of the categorization of underlying latent variables, (iii) non-negative half student-t priors with the same parameter values as the above were employed for the standard deviation of random effects, and (iv) the LKJ distribution was specified as a prior for the aforementioned correlations of random intercepts. The posterior distribution was derived based on Hamiltonian Monte Carlo with four Markov chains with 10,000 iterations each, including 2,000 warmup iterations.

R-hat indices were all below 1.01, which suggested model convergence. Full posterior distributions are shown in **Supporting information-F**. In order to assess the goodness of fit of the model, the ratings with the highest posterior probabilities and the observed ratings were cross-tabulated. Out of the 2,200 ratings (i.e., 110 participants × 10 raters × 2 outcome measures), the model classified 763 ratings (34.7%) squarely into one of the nine categories. This, however, could be due to random intercepts. In order to isolate the effects of factor scores from random effects, we rebuilt the model that only included 11 factors, and compared its classification accuracy with the baseline accuracy, where we classified all the ratings into the largest category in each outcome variable (i.e., 204 ratings in Rating = 7 in comprehensibility and 179 ratings in Rating = 4 in nativelikeness). The difference in classification accuracy between the two reflects the effects of the 11 factors. The classification accuracy based on the model with 11 factors was 456 (20.7%), whereas the baseline accuracy was 383 (17.4%). The difference between the two ratios was significant ($\chi^2(1) = 5.16$, $p = .023$). Although the extra accuracy brought by the 11 factors might not look

large, it is arguably still acceptable considering that much of the variability in ratings stems from the learner-rater interaction. This is exemplified by the fact that the classification accuracy is merely 35% even when between-learner and between-rater variability is perfectly accounted for by random effects, and the main predictors of the model are factor scores that do not explain the interaction. Furthermore, if an error by one rating is allowed (e.g., a speech sample that received the rating of 5 and was misclassified as a 6 is counted as an instance of accurate classification), then the accuracy rises to 1,156 ratings (52.5%), with the baseline accuracy of 1,060 ratings (48.2%). Therefore, the model fits the data reasonably well, and the inferences based on the model are considered to be credible.

Table 1

Summary of the Bayesian Multivariate Mixed-Effects Ordinal Regression Model

Parameter	Comprehensibility			Nativelikeness		
	Mean	95% Credible Interval		Mean	95% Credible Interval	
		Lower	Upper		Lower	Upper
Threshold (1 vs 2)	-5.94	-7.42	-4.53	-4.05	-4.98	-3.11
Threshold (2 vs 3)	-4.60	-6.03	-3.23	-2.50	-3.42	-1.59
Threshold (3 vs 4)	-3.39	-4.79	-2.04	-1.14	-2.04	-0.23
Threshold (4 vs 5)	-2.33	-3.73	-0.98	0.09	-0.80	1.00
Threshold (5 vs 6)	-1.31	-2.70	0.05	1.17	0.27	2.09
Threshold (6 vs 7)	-0.22	-1.60	1.12	2.48	1.57	3.41
Threshold (7 vs 8)	1.26	-0.13	2.61	3.72	2.79	4.66
Threshold (8 vs 9)	3.15	1.75	4.52	5.51	4.53	6.52
Factor 1: Experience Quantity	0.09	-0.24	0.42	0.15	-0.25	0.54
Factor 2: Current L2 Use	0.43	0.10	0.76	0.55	0.15	0.94
Factor 3: Awareness of Nativeness	-0.01	-0.33	0.32	0.06	-0.32	0.44
Factor 4: Age of Immersion	-0.44	-0.75	-0.11	-0.64	-1.02	-0.25
Factor 5: Motivation	0.08	-0.24	0.41	-0.01	-0.39	0.38
Factor 6: Attitude to Nativeness	0.39	0.07	0.72	0.41	0.03	0.80
Factor 7: EFL Experience	-0.17	-0.49	0.16	-0.31	-0.69	0.08
Factor 8: Special Past Experience	-0.44	-0.77	-0.11	-0.44	-0.83	-0.05
Factor 9: Attitude to Foreign Accents	0.01	-0.32	0.34	0.02	-0.37	0.42
Factor 10: Comprehensibility Orientation	0.05	-0.26	0.37	-0.04	-0.41	0.33
Factor 11: L1 Influence	-0.13	-0.46	0.20	-0.42	-0.80	-0.03
SD of by-learner random intercepts	1.60	1.35	1.90	1.95	1.65	2.29
SD of by-rater random intercepts	1.96	1.16	3.46	1.26	0.74	2.25
Correlation between two outcome measures						
By-learner random intercepts	0.98	0.95	1.00			
By-rater random intercepts	-0.19	-0.73	0.46			

Table 1 shows the posterior mean and the 95% credible intervals (central posterior intervals) of each parameter. The threshold parameters represent the threshold values of

categorization of the continuous latent variable assumed to underlie the ordinal outcome variable. Our focal interest concerns Factors 1 through 11. Since both latent variables underlying outcome variables and factor scores are in unit scale, the parameter values indicate the change in the latent variable in standard deviation (*SD*) associated with one *SD* change in factor scores. The table shows that in both comprehensibility and nativelikeness ratings, zero fell outside of the credible intervals (*CIs*) in Factors 2 (Current L2 Use), 4 (Age of Immersion), 6 (Attitude to Nativeness), and 8 (Special Past Experience). Additionally, the CIs of Factor 11 (L1 Influence) did not include the null effect in nativelikeness ratings. Factors 2 and 6 are positively correlated with higher ratings, while Factors 4 and 8 are negatively correlated in both outcome measures. Factor 11 (L1 Influence) is negatively correlated with nativelikeness ratings (but not comprehensibility ratings).

The results of Bayesian analyses are influenced by the choice of prior distributions. In order to investigate the potential effects of priors, we rebuilt the model with different priors for slope parameters, which are the focus of this study. Specifically, we gradually increased the standard deviation of the normal distribution from 0.8 to 3, and also tested a flat prior. The results largely remained the same. The details are reported in **Supporting Information-G**.

Varying Strengths Across Ratings

We conducted an additional analysis on the extent to which the strength of the five prominent factors (Current L2 Use, Age of Immersion, Nativeness Attitude, Special Past Experience, and L1 Influence) would differ depending on different levels of L2 comprehensibility and nativelikeness. See **Supporting Information-H**.

Discussion

Despite much scholarly discussion directed toward the sources of individual differences in L2 speech learning in adulthood, the transparency, size, and diversity of datasets in prior work have remained problematic. To move ahead the research agenda, in our novel study, we took the dynamic perspective on L2 learning (including multiple independent and dependent variables; Larsen-Freeman, 2012) and the Open Science approach (making the details of our own dataset publicly available; Marsden, in press). Specifically, we first presented the dataset of speech samples and the questionnaires from 110 late L2 learners in London. Subsequently, we demonstrated the way we expounded the complex relationship

between a total of 30 variables of learner-external and -internal individual differences—L1 backgrounds, age, experience, motivation, awareness and attitudes—and two different dimensions of L2 speech proficiency—comprehensibility and nativelikeness. As reviewed earlier, the existing literature has found all the learner variables selected for this study to affect L2 speech proficiency to some degree. The primary objective of the current investigation was to reveal the *relative* weights of these variables by way of mixed effects modeling analyses.

According to the results of the analyses, these between-learner variables allowed 20.7% of the ratings to be classified accurately, which we consider robust and comparable to previous research using similar mixed effects models. Among all the associated variables, it was five factors that showed particularly observable associations—i.e., current, past and special experience, attitude and L1-L2 distance. In essence, L2 learners who have received higher comprehensibility scores, and by extension have achieved higher L2 speech proficiency levels, use L2 English on a regular basis. These L2 users interact more often with fluent (rather than non-fluent) speakers in L2 English (rather than their L1) (i.e., current experience factors). Not only have these learners arrived in an L2 speaking environment in early adulthood, entailing longer length of immersion (i.e., age factors), but also have had extra, professional experience related to linguistic training and L2 English teaching (i.e., special experience factors). Finally, these learners tend to engage in every L2-use related opportunity with a more positive attitude toward the language of the community, i.e. British English (i.e., learner-internal, attitude factors). To achieve more *nativelike* L2 speech, however, the results indicated that L1-L2 distance may play a significant role. In the case of our study, those who spoke an Indo-European language as an L1 likely showed less detectable L2 accent and thus attained more nativelike oral proficiency (i.e., L1 influence factors).

Assuming that L2 speech proficiency develops over time on the continuum from low to advanced, the results of our cross-sectional dataset provided empirical support to the view that the comprehensibility and nativelikeness aspects of L2 speech learning are comprised of slightly different processes. L2 comprehensibility development continues to take place during adulthood, as long as learners frequently practice a target language in various social settings (Derwing & Munro, 2013; Saito et al., 2017) with positive attitude and orientation toward the target language and its community (Dewaele, Whitney, Saito, & Dewaele, 2018). Although many L2 learners strive to approximate the nativelike aspects of L2 speech, foreign accent

reduction seems to be tied to factors that most learners cannot control on their own: Attaining more nativelike L2 pronunciation may be limited to certain individuals whose L1-L2 distance is relatively small (i.e., other Indo-European languets) (Bongaerts et al., 1997).

Taken together, the findings support an increasingly popular idea that L2 learning is a dynamic, complex, adaptive system within which a range of learner external and internal factors affect each other (e.g., Larsen-Freeman, 2012). Following this line of thought, we argue that it is crucial for future L2 speech research to include multiple affecting factors related to contexts and individuals instead of examining each single variable in isolation. To tackle the topic of individual differences in any aspect of L2 learning, much caution needs to be exercised in data collection and analysis. It is important to recruit a large number of participants to maintain a strong statistical power of dependent variables, minimize the number of independent variables via data reduction (e.g., factor analyses), and inspect the dynamic, complex link between dependent and independent variables via Bayesian multivariate mixed-effects analyses.

Although we believe our statistical data analysis is reasonable, it is certainly not the only valid way to analyze our data (see **DATASET**). In psychology, different analyses of a single dataset have been demonstrated to yield different results even for the same research question (Silberzahn et al., 2018). Thus, we welcome any interested readers to reanalyze our data in the way they prefer and examine any potential differences that arise between their results and ours. Together with such future analyses, we hope to collectively realize a multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016), in which a single raw dataset is analyzed in a variety of ways to gain insights into how much results may change due to the (arbitrary) decisions researchers make during their data analysis (i.e., so-called ‘researcher degrees of freedom’; Simmons, Nelson, & Simonsohn, 2011).

Below, we offer a few alternative, arguably equally valid means by which to analyze our dataset.

1. While we employed a factor score regression (i.e., a factor analysis followed by a regression analysis using the factor scores), one could also build a single structural equation model (SEM) that encompasses both factor and regression models. The SEM can presumably better propagate uncertainty from a measurement model (corresponding to the factor analysis) to a structural model (corresponding to the regression analysis).

2. Another approach is to use penalized regressions without relying on a factor analysis to reduce the number of predictors. Common penalized regression methods such as lasso regression and ridge regression can be viewed as regression models with regularizing prior probabilities on parameter values in a Bayesian sense. Since variables are not reduced, interpretations might turn out to be less challenging with this approach.
3. Furthermore, one could also view the analytical task as one of classification and employ machine learning techniques to predict the ratings of speech samples based on the combination of variables available, after which they could examine which variables influenced the classification.
4. Finally, one can also perform the frequentist analysis equivalent to the Bayesian analyses we performed, and examine whether the results converge.

References

Best, C., & Tyler, M. (2007). Nonnative and second-language speech perception. In O. Bohn, & M. Munro (Eds.), *Language experience in second language speech learning: In honour of James Emil Flege* (pp. 13–34). Amsterdam: John Benjamins.

Boersma, P., & Weenink, D. (2017). PRAAT: doing phonetics by computer [Computer program]. Version 6.0.28.

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, *19*(4), 447-465.

Bürkner, P. (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software, 80*, 1-28.

Bürkner, P. C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science, 2*(1), 77-101.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1-32. doi: 10.18637/jss.v076.i01.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, *63*(2), 163-185.

Dewaele, J. M., Witney, J., Saito, K. & Dewaele, L. (2018). Foreign language enjoyment and anxiety in the FL classroom: The effect of teacher and learner variables. *Language Teaching Research, 22*(6), 676-697. doi: 10.1177/1362168817692161

Elliott, A. R. (1995). Field independence/dependence, hemispheric specialization, and attitude in relation to pronunciation accuracy in Spanish as a foreign language. *The Modern Language Journal*, *79*(3), 356-371.

Flege, J. E., Munro, M. J., & MacKay, I. R. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, *97*(5), 3125-3134.

Flege, J. E., MacKay, I. R., & Piske, T. (2002). Assessing bilingual dominance. *Applied Psycholinguistics*, *23*(4), 567-598.

Freed, B., Dewey, D., Segalowitz, N., & Halter, R. (2004). The language contact profile. *Studies in Second Language Acquisition*, *26*(2), 349-356.

Gardner, R.C., & Smythe, P.C. (1981). On the development of the Attitude/Motivation Test Battery. *Canadian Modern Language Review, 37*(3), 510-525.

Garrett, P. (2010). *Attitudes to language*. Cambridge: Cambridge University Press.

Gatbonton, E., & Trofimovich, P. (2008). The ethnic group affiliation and L2 proficiency link: Empirical evidence. *Language Awareness*, *17*(3), 229-248.

Gewin V. (2016). Data sharing: An open mind on open data. *Nature, 529,* 117-119.

Gerrig, R., & Rastle, K. (2019). New initiatives to promote open science at the Journal of Memory and Language. *Journal of Memory and Language*, *104,* 126-127.

Hama, M., & Leow, R. P. (2010). Learning without awareness revisited: Extending Williams (2005). *Studies in Second Language Acquisition, 32*(3), 465-491.

Harrell, F. E., Jr. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis* (2nd edition). New York: Springer.

Hopp, H., & Schmid, M. S. (2013). Perceived foreign accent in first language attrition and second language acquisition: The impact of age of acquisition and bilingualism. *Applied Psycholinguistics*, *34*(2), 361-394.

Hui, B. (2019). Analyzing processing time data in applied linguistics and second language research: A multivariate mixed-effects approach. Journal of Research Design and Statistics in Linguistics and Communication Science, 5(1-2), 189-207. https://doi.org/10.1558/jrds.39117

Ioup, G., Boustagi, E., El Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis: A case study of successful adult SLA in a naturalistic environment. *Studies in Second Language Acquisition*, *16*(1), 73-98.

Jia, G., & Aaronson, D. (2003). A longitudinal study of Chinese children and adolescents learning English in the United States. *Applied Psycholinguistics*, *24*(1), 131-161.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd edition). London: Academic Press.

Lambert, B. (2018). *A student's guide to Bayesian statistics*. London: Sage Publications.

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*(S1)*,* 127-159.

Larsen-Freeman, D. (2012). Complex, dynamic systems: A new transdisciplinary theme for applied linguistics? *Language Teaching*, *45*(2), 202-214.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328-348.

Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In R. Plonsky (Ed.). *Advancing quantitative methods in second language research* (pp. 182-212). New York: Routledge.

Marsden, E. (in press). Open science in applied linguistics and its consequences for quality and scope of research. In J. McKinley and H. Rose (Eds.), *Routledge handbook of research methods in applied linguistics.* New York: Routledge.

Marsden, E., Crossley, S., Ellis, N., Kormos, J., Morgan-Short, K., & Thierry, G. (2019). Inclusion of Research Materials When Submitting an Article to Language Learning. *Language Learning*, *69*(4), 795-801.

Marsden, E., Mackey, A., & Plonsky, L. D. (2016). The IRIS repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1-21). New York: Routledge.

Marsden, E., Morgan-Short, K., Thompson, S. & Abugaber, D. (2018). Replication in second language research: narrative and systematic reviews and recommendations for the field. *Language Learning, 68*(2), 321-391.

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*(5), 861-904. doi:10.1017/S0142716418000036

McKenzie, R. M., & Gilmore, A. (2017). "The people who are out of 'right' English": Japanese university students' social evaluations of English language diversity and the internationalisation of Japanese higher education. *International Journal of Applied Linguistics*, *27*(1), 152-175.

McKiernan E. C., Bourne P. E., Brown C. T., Buck S., Kenall A., Lin J., Yarkoni T., et al (2016). How open science helps researchers succeed. *eLife, 5*, e16800.

Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics, 35*, 463–482.

Muñoz, C. & Llanes, À. (2014). Study abroad and changes in degree of foreign accent in children and adults. *The Modern Language Journal*, *98*(1), 432-449.

Munro, M. (2013). *What do you know when you "know" an L2 vowel?* In Paper presented at Pronunciation in Second Language Learning and Teaching Conference, Ames, IA.

Norouzian, R., de Miranda, M., & Plonsky, L. (2018). The Bayesian revolution in second language research: An applied approach. *Language Learning*, *68*(4), 1032-1075.

Ortega, L. (2018). Ontologies of language, second language acquisition, and world Englishes. *World Englishes*, *37*(1), 64-79.

Piske, T., Flege, J., MacKay, & Meador, D. (2011). Investigating native and non-native vowels produced in conversational speech. In M. Wrembel, M. Kul & K. Dziubalska-Kołaczyk (Eds.), *Achievements and perspectives in the acquisition of second language speech: New Sounds 2010* (pp. 195–205). Frankfurt am Main: Peter Lang.

Plonsky, L., Marsden, E. J., Crowther, D., Gass, S., & Spinner, P. (2019). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research*, 1-39.

Saito, K. (2013). Age effects on late bilingualism: The production development of /r/ by high-proficiency Japanese learners of English. *Journal of Memory and Language, 69*, 546-562.

Saito, K. (2019). Individual differences in second language speech learning in classroom settings: Roles of awareness in the longitudinal development of Japanese learners' English /ɹ/ pronunciation. *Second Language Research, 35*(2), 149-172.

Saito, K., Dewaele, J.-M., Abe, M., & In'nami, Y. (2018). Motivation, emotion, learning experience and second language comprehensibility development in classroom settings: A cross-sectional and longitudinal study. *Language Learning, 68*, 709-743.

Saito, K., & Munro, M. (2014). The early phase of /r/ production development in adult Japanese learners of English. *Language and Speech, 57*(4), 451-469.

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning, 69(3),* 652-708.

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

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*(4), 439-462.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 1–32). Cambridge, UK: Cambridge University Press.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., & Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 3*, 337-356.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702-712.

Suzuki, S. & Kormos, J. (2019). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*. 1-25. doi:10.1017/S0272263119000627

Venkatagiri, H. S., & Levis, J. M. (2007). Phonological awareness and speech comprehensibility: An exploratory study. *Language Awareness*, *16*(4), 263-277.

Wang, Y., Jongman, A., & Sereno, J. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustical Society of America,* 113, 1033–1043.

SUPPORTING INFORMATION-A: SPEAKER QUESTIONNAIRE

Basic info

(1). Age: years old
(2). Age of Arrival: years old
(3). Where? ① UK ② North America ③ Australia/NZ ④ Others ( )
(4). Why? ① Study abroad ② Work abroad ③ Immigration ④ Others ( )
(5). Have you ever taught English before? years (e.g., 0-10 years)
(6). Have you taken any linguistics classes/training before? (0 = no; 1 = yes)

Length of residence

(7). UK: years
(8). North America: years
(9). Australia/NZ: years
(10). Others: years (0 = NO) (which countries?)

L2 English Learning in Classroom Settings

(11). Age of learning in classroom settings: years old
(12). Length of learning in classroom settings: years

First Language

(13). First language (from birth/most dominant): Which language? ( )

L2 English Use Profile: Average over the past 1-2 years

Frequency at work/school (professional settings)
(14). L1/most dominant: % (0-100%)
(15). English (with fluent speakers): % (0-100%)
(16). English (with non-fluent speakers): % (0-100%)
Frequency with friends (social settings)
(17). L1/most dominant: % (0-100%)
(18). English (with fluent speakers): % (0-100%)
(19). English (with non-fluent speakers): % (0-100%)
Frequency at home
(20). L1/most dominant % (0-100%)
(21). English (with fluent speakers): % (0-100%)
(22). English (with non-fluent speakers): % (0-100%)

Motivation and Concern for "Nativelikeness" (RP)

To what degree are you expected to use L2 English at a nativelike proficiency level on a 9-point scale (1 = not at all, 9 = very much so)?

(23). At work/school (professional settings)
(24). With friends (social settings)
(25). At home

Awareness of one's own L2 English oral proficiency

While speaking L2 English, which aspects of language do you think are relatively crucial for successful communication? Please rate the following statements on a 9-point scale (1 = not important, 9 = very important)?

(26). Speaking English without any accent like a native speaker
(27). Speaking comprehensible English regardless of accentedness
(28). Good pronunciation
(29). Appropriate vocabulary/grammar
(30). Idiomatic & sophisticated expression

Perception of Foreign Accents

How much do you agree with the following statements (1 = strongly disagree, 9 = strongly agree)?

(31). I like it when people speak English with a foreign accent
(32). I like it when people speak English with British accent (RP)
(33). I like it when people speak English with American accent (GA).

How much are you familiar with different types of English (1 = I am not familiar at all, 9 = I am very much)

(34). I am familiar with British English (Received Pronunciation) (1-9)
(35). I am familiar with different kinds of foreign accented English (1-9)

SUPPORTING INFORMATION-B: Descriptive Statistics of 30 Learner Variables Among 110 L2 Participants

	M	*SD*	*Range*	
			Min	*Max*
First language				
L1-L2 distance	Indo-European (46), Non Indo-European (64)			
A. Age				
Chronological age	30.4	7.4	20	59
Age of arrival	24.4	6.0	16	55
Age of foreign language learning	9.9	5.8	2	58
B. Previous Experience				
Length of residence in English speaking environments	4.9	6.2	0.4	39
Length of residence in the UK	4.4	6.0	0.1	39
Length of foreign language learning	11.5	4.8	0	23
Previous linguistics training experience	Yes (33), No (77)			
Previous English teaching experience	Yes (31), No (79)			
C. Current Experience				
L1 use at work	15.1%	22.1	0	80
L2 use with fluent speakers at work	68.1%	29.4	10	100
L2 use with non-fluent speakers at work	15.0%	19.8	0	88
L1 use in social settings	40.2%	28.7	0	100
L2 use with fluent speakers in social settings	48.4%	29.2	0	100
L2 use with non-fluent speakers in social settings	8.4%	14.2	0	80
L1 use at home	60.4%	39.3	0	100
L2 use with fluent speakers at home	32.9%	38.5	0	100
L2 use with non-fluent speakers at home	2.8%	8.9	0	60
D. Motivation				
Expectation at work	6.1	2.3	1	9
Expectation in social settings	4.6	2.4	1	9
Expectation at home	3.0	2.5	1	9
E. Awareness				
Awareness of nativelikeness	5.0	2.3	1	9
Awareness of comprehensibility	8.2	1.0	4	9
Awareness of pronunciation accuracy	7.3	1.5	2	9
Awareness of appropriate lexicogrammar	6.9	1.7	1	9
Awareness of idiomatic expression	4.6	2.2	1	9
F. Familiarity and Attitude				
Familiarity towards foreign accent	5.0	2.3	1	9
Familiarity towards British English	8.2	1.0	4	9
Attitude towards foreign accent	7.3	1.5	2	9
Attitude towards British English	6.9	1.7	1	9

SUPPORTING INFORMATION-C: Training Scripts and Onscreen Labels of L2 Comprehensibility and Nativelikeness Judgements

A. Training scripts for comprehensibility and nativelikeness judgement

Comprehensibility	This term refers to how much effort it takes to understand what someone is saying. If you can understand with ease, then a speaker is highly comprehensible. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility.
Nativelikeness	This refers to how much a second language speech sample differs from the variety of English commonly used in the community.

B. Onscreen labels

Comprehensibility								
1	2	3	4	5	6	7	8	9
Hard to understand							Easy to understand	
Nativelikeness								
1	2	3	4	5	6	7	8	9
Not nativelike at all							Very nativelike	

SUPPORTING INFORMATION-D: Summary of a Ten-Factor Solution Based on a Factor Analysis of the Learner Background Questionnaire

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Factor 11
Labels	Experience Quantity	Current L2 Use	Awareness of Nativeness	Age of Immersion	Motivation	Attitude to Nativeness	EFL Experience	Special Past Experience	Foreign Accents Attitude	Comprehensibility Orientation	L1 Influence
Cumulative %	15.39	27.14	36.82	46.07	54.07	59.88	65.29	70.03	74.59	78.76	82.55
First language											
L1-L2 distance	-0.019	-0.090	0.135	-0.018	0.032	-0.092	0.110	-0.055	0.180	-0.015	**0.902**
A. Age											
Chronological age	**0.610**	0.027	0.010	**0.685**	-0.003	-0.002	-0.132	-0.030	0.054	0.001	-0.046
Age of arrival	-0.186	-0.044	-0.009	**0.924**	-0.032	0.057	-0.026	-0.016	-0.065	-0.012	-0.025
Age of foreign language learning	-0.048	0.014	0.073	0.264	-0.037	-0.051	**-0.851**	-0.030	-0.056	0.034	0.020
B. Previous Experience											
Length of residence in English speaking environments	**0.973**	0.015	-0.010	-0.091	0.019	-0.016	-0.041	-0.009	0.019	-0.027	0.013
Length of residence in the UK	**0.965**	0.017	-0.010	-0.037	0.016	-0.008	-0.019	0.027	0.013	-0.058	-0.010
Length of foreign language learning	-0.160	0.129	0.000	0.221	-0.097	-0.019	**0.836**	0.010	-0.060	0.023	0.120
Previous linguistics training experience	-0.069	0.042	0.055	-0.041	0.030	-0.012	-0.104	**-0.893**	0.079	-0.046	0.094
Previous English teaching experience	0.070	0.058	-0.099	0.022	-0.062	0.050	0.054	**-0.926**	-0.100	0.090	-0.047
C. Current Experience											
L1 use	0.017	**-0.901**	0.019	-0.016	0.032	-0.040	-0.046	0.066	-0.057	-0.058	0.026
L2 use with fluent speakers	0.035	**0.928**	0.081	-0.031	0.008	-0.040	0.029	-0.049	-0.040	-0.001	-0.021
D. Motivation											
Expectation at work	0.000	-0.124	-0.026	-0.084	**0.825**	0.078	0.038	-0.039	-0.084	0.277	-0.063
Expectation in social settings	0.031	-0.011	0.024	-0.059	**0.889**	0.044	-0.093	0.077	-0.035	0.029	0.110
Expectation at home	0.011	0.157	-0.016	0.184	**0.649**	-0.033	0.037	-0.062	0.193	-0.376	-0.058
E. Awareness											
Awareness of nativelikeness	-0.104	0.026	**0.874**	-0.100	-0.044	0.166	-0.096	-0.038	-0.062	-0.078	0.078
Awareness of comprehensibility	-0.125	0.020	0.067	0.046	0.166	-0.217	0.068	-0.126	-0.032	**0.760**	0.035
Awareness of pronunciation accuracy	0.167	0.022	**0.668**	0.171	0.080	0.084	0.087	0.075	-0.134	0.239	0.158
Awareness of appropriate lexicogrammar	0.085	-0.232	0.351	0.129	-0.004	0.058	0.164	-0.174	0.373	0.088	-0.435
Awareness of idiomatic expression	0.008	0.070	**0.610**	-0.001	0.054	-0.282	-0.025	0.019	0.184	-0.068	-0.426
F. Familiarity and Attitude											
Familiarity towards foreign accent	0.007	0.191	-0.051	-0.087	-0.077	0.264	-0.092	0.143	0.318	**0.605**	-0.150

Familiarity towards British English	0.180	0.099	0.110	-0.196	0.086	**0.721**	0.173	-0.046	0.022	-0.015	-0.116
Attitude towards foreign accent	0.041	0.050	-0.099	-0.039	0.000	0.008	0.000	0.009	**0.927**	0.031	0.161
Attitude towards British English	-0.124	-0.039	0.064	0.174	0.045	**0.868**	-0.064	0.015	0.011	-0.075	0.000

SUPPORTING INFORMATION-E: Visual Inspection of Relationships Between L2 Speech Ratings and Learner Factors

Below, we can see that there is some systematicity between factor scores and the ratings in some factors. Specifically, in both outcome measures, Factors 2, 6, and, to a lesser extent, 9 appear to be positively correlated with ratings while Factors 4, 8, and 11 appear to be negatively correlated. Variability of factor scores, however, tends to be large within each rating in each factor, and no firm conclusion can be drawn from the figure alone. We, therefore, tested the relationship in a more formal manner.

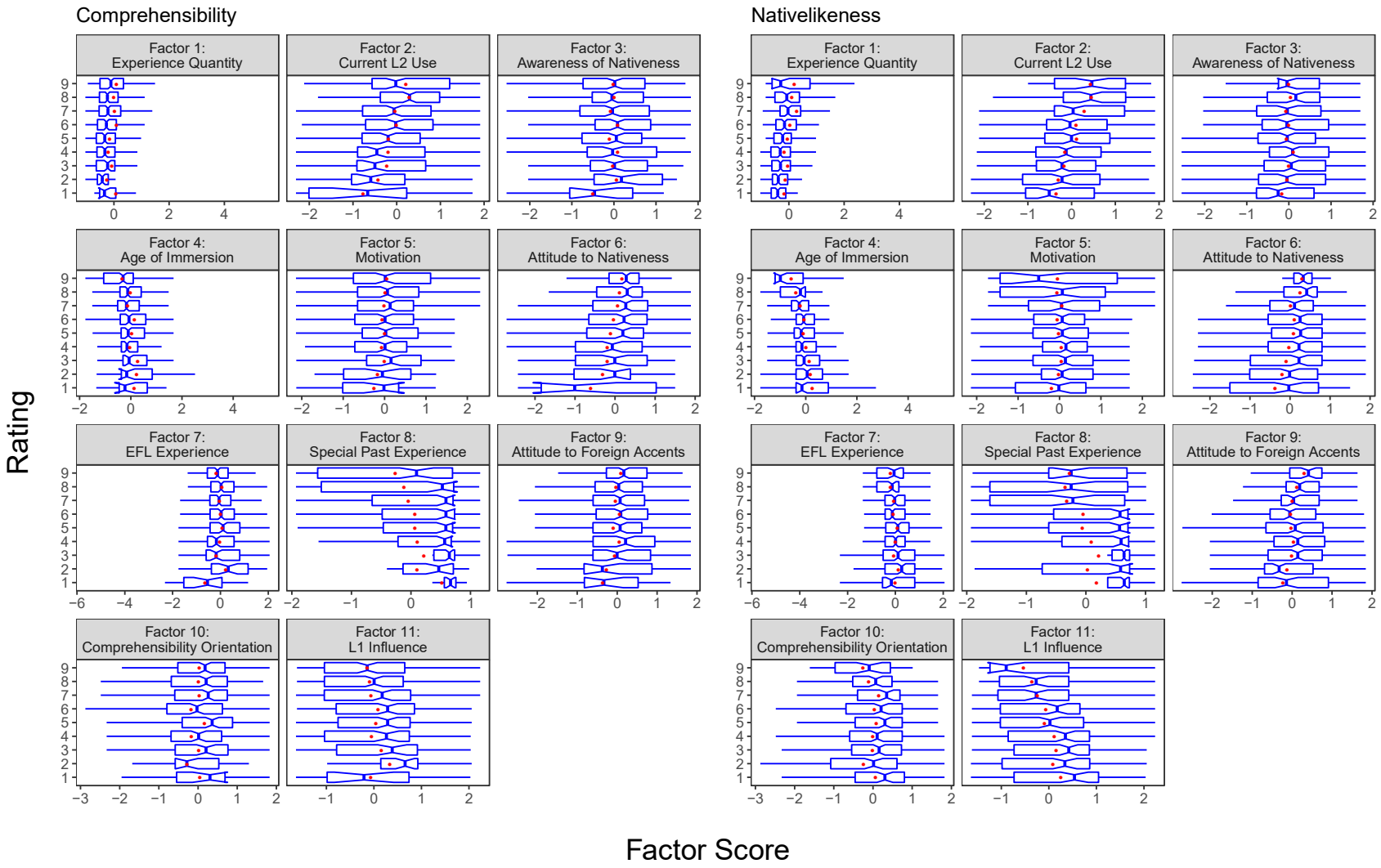

Figure 1. Distribution of factor scores in each rating of comprehensibility and nativelikeness in each factor. Grey dots represent individual judgments, while a larger dot represents the mean value in each rating category.

SUPPORTING INFORMATION-G: Posterior Distribution

Figures 1 through 3 show the posterior distribution of the population-level parameters, SDs of random intercepts, and correlation parameters in the Bayesian multivariate mixed-effects ordinal regression model. Most of the distributions appear to be reasonably normal, and even when they deviate from a normal distribution, the CIs seem summarise the distribution to a good extent (e.g., SDs of random intercepts in Figure 3). The only possible exception is the correlation of by-participant random intercepts (i.e., lower-mid panel in Figure 3), which appears to have reached the ceiling and the mean of the distribution differs from its mode. The interpretation of the parameter, therefore, requires some caution. We do not interpret it.

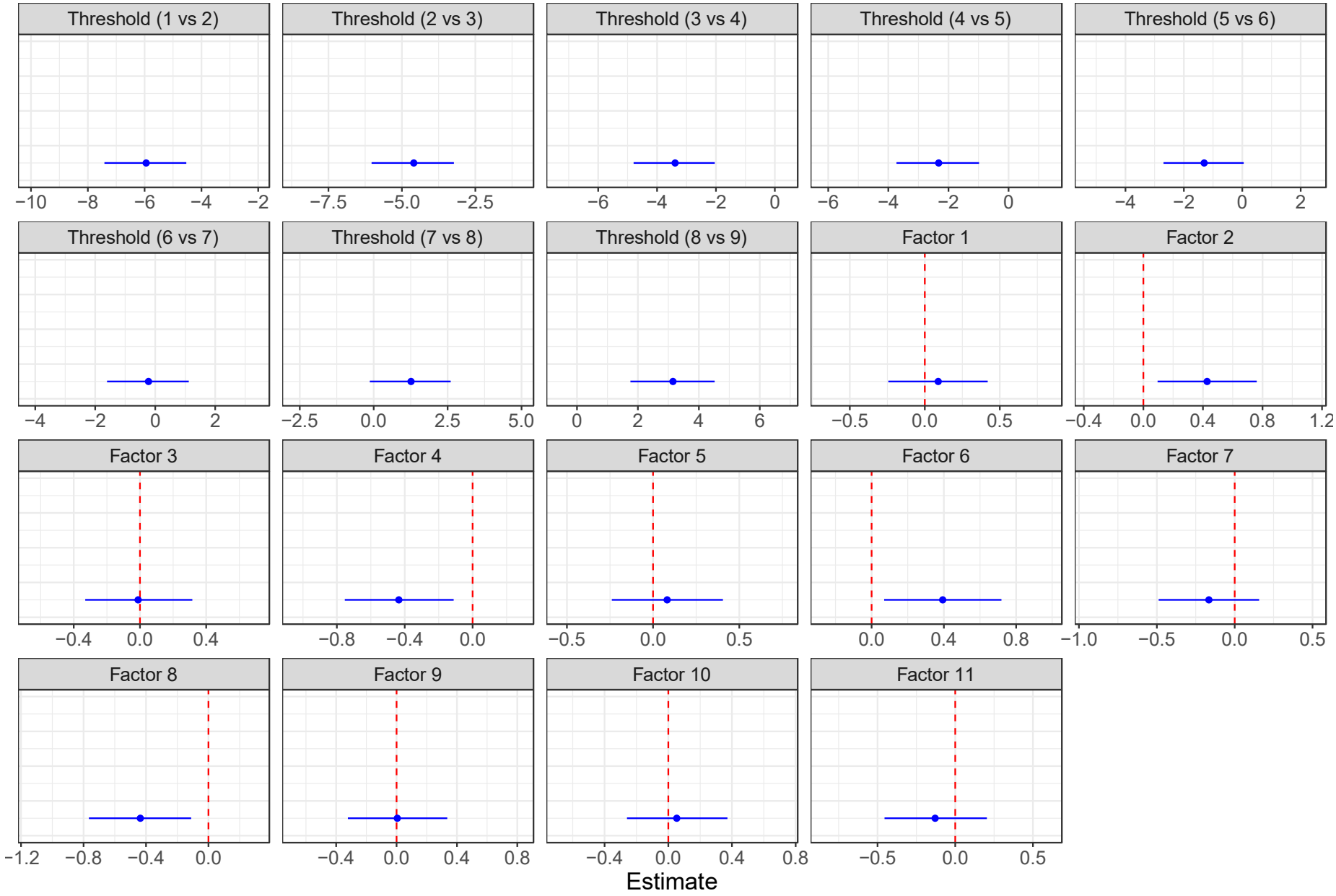
Threshold (1 vs 2)
Threshold (2 vs 3)
Threshold (3 vs 4)
Threshold (4 vs 5)
Threshold (5 vs 6)
Threshold (6 vs 7)
Threshold (7 vs 8)
Threshold (8 vs 9)
Factor 1
Factor 2
Factor 3
Factor 4
Factor 5
Factor 6
Factor 7
Factor 8
Factor 9
Factor 10
Factor 11
Estimate

Figure 1. Posterior distribution of the fixed-effects parameters in comprehensibility ratings. Blue horizontal lines represent 95% CIs, while red vertical lines (only drawn for slope parameters) represent the null effect.

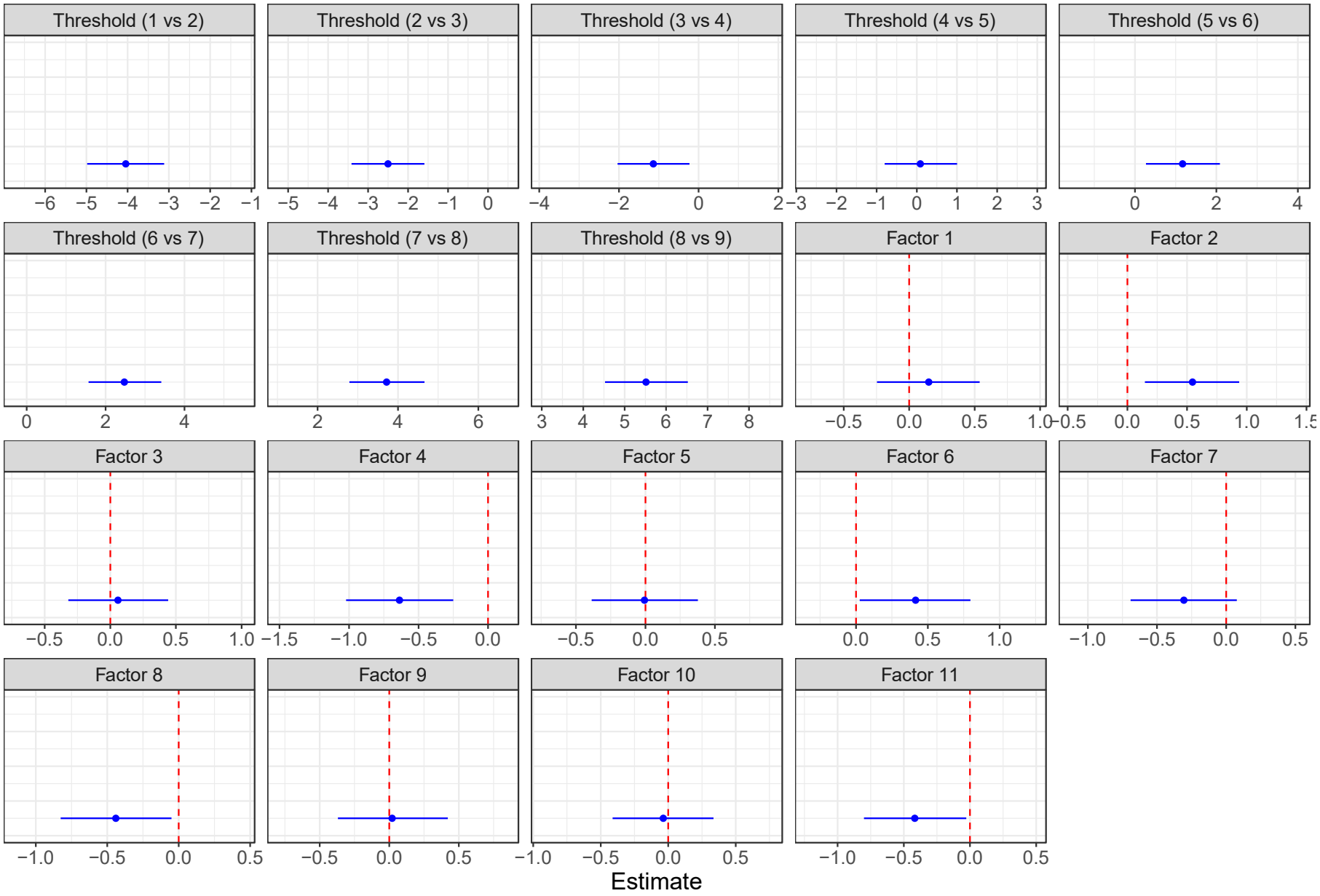
Threshold (1 vs 2)
Threshold (2 vs 3)
Threshold (3 vs 4)
Threshold (4 vs 5)
Threshold (5 vs 6)
Threshold (6 vs 7)
Threshold (7 vs 8)
Threshold (8 vs 9)
Factor 1
Factor 2
Factor 3
Factor 4
Factor 5
Factor 6
Factor 7
Factor 8
Factor 9
Factor 10
Factor 11
Estimate

Figure 2. Posterior distribution of the fixed-effects parameters in nativelikeness ratings. See the caption of Figure 1 for the interpretation.

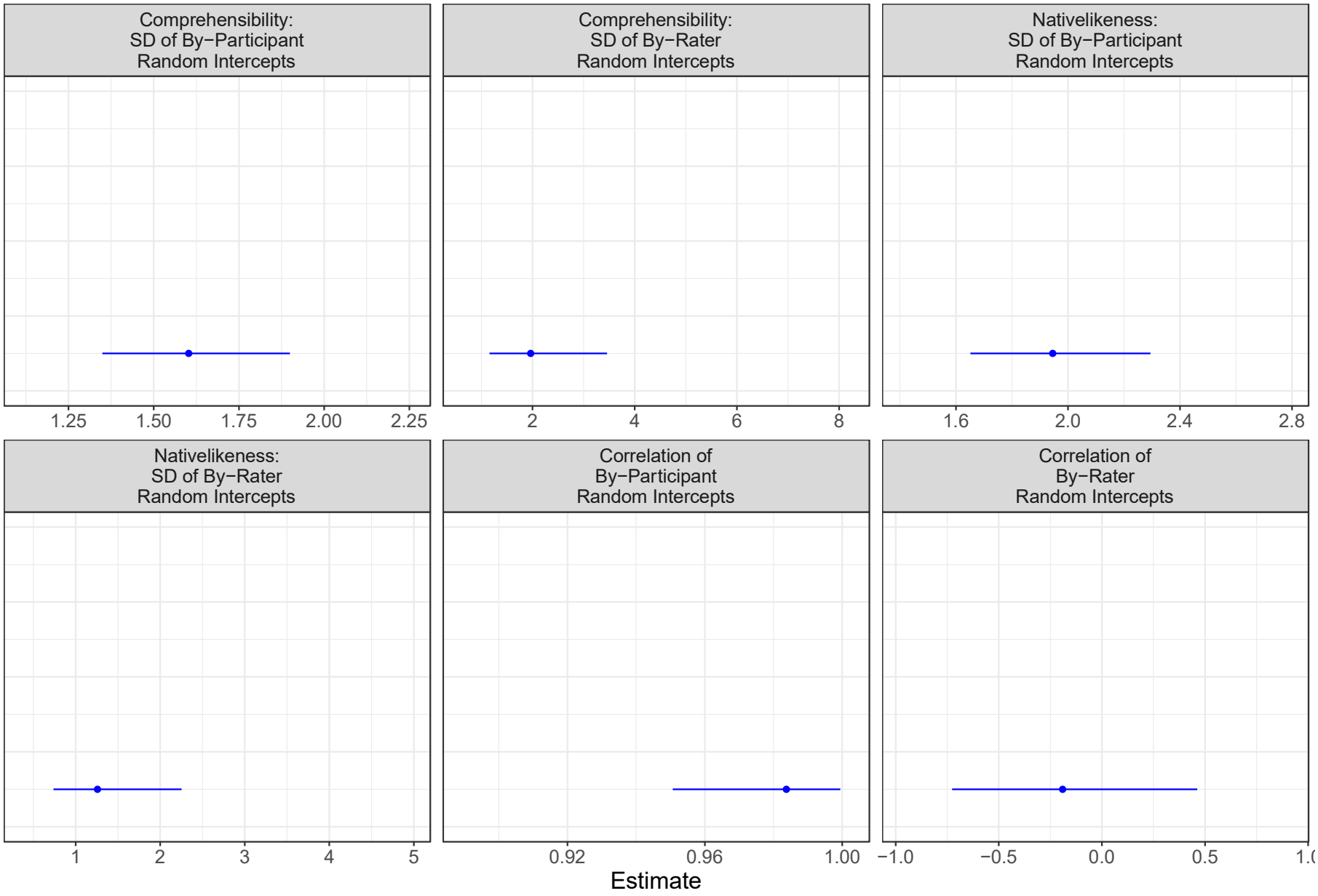
Comprehensibility:
SD of By−Participant
Random Intercepts
1.25
1.50
1.75
2.00
2.25
Comprehensibility:
SD of By−Rater
Random Intercepts
2
4
6
8
Nativelikeness:
SD of By−Participant
Random Intercepts
1.6
2.0
2.4
2.8
Nativelikeness:
SD of By−Rater
Random Intercepts
1
2
3
4
5
Correlation of
By−Participant
Random Intercepts
0.92
0.96
1.00
Correlation of
By−Rater
Random Intercepts
−1.0
−0.5
0.0
0.5
Estimate

Figure 3. Posterior distribution of random intercepts and correlation parameters. See the caption of Figure 1 for the interpretation.

SUPPORTING INFORMATION-H: Sensitivity Analysis

Figures 1 and 2 show the posterior means and their 95% CIs across different prior distributions in comprehensibility ratings and nativelikeness ratings, respectively. We can observe that neither the point estimate of the parameter nor its uncertainty is affected much by the choice of the priors considered here. The inferences drawn, therefore, is considered fairly robust against the choice of priors.

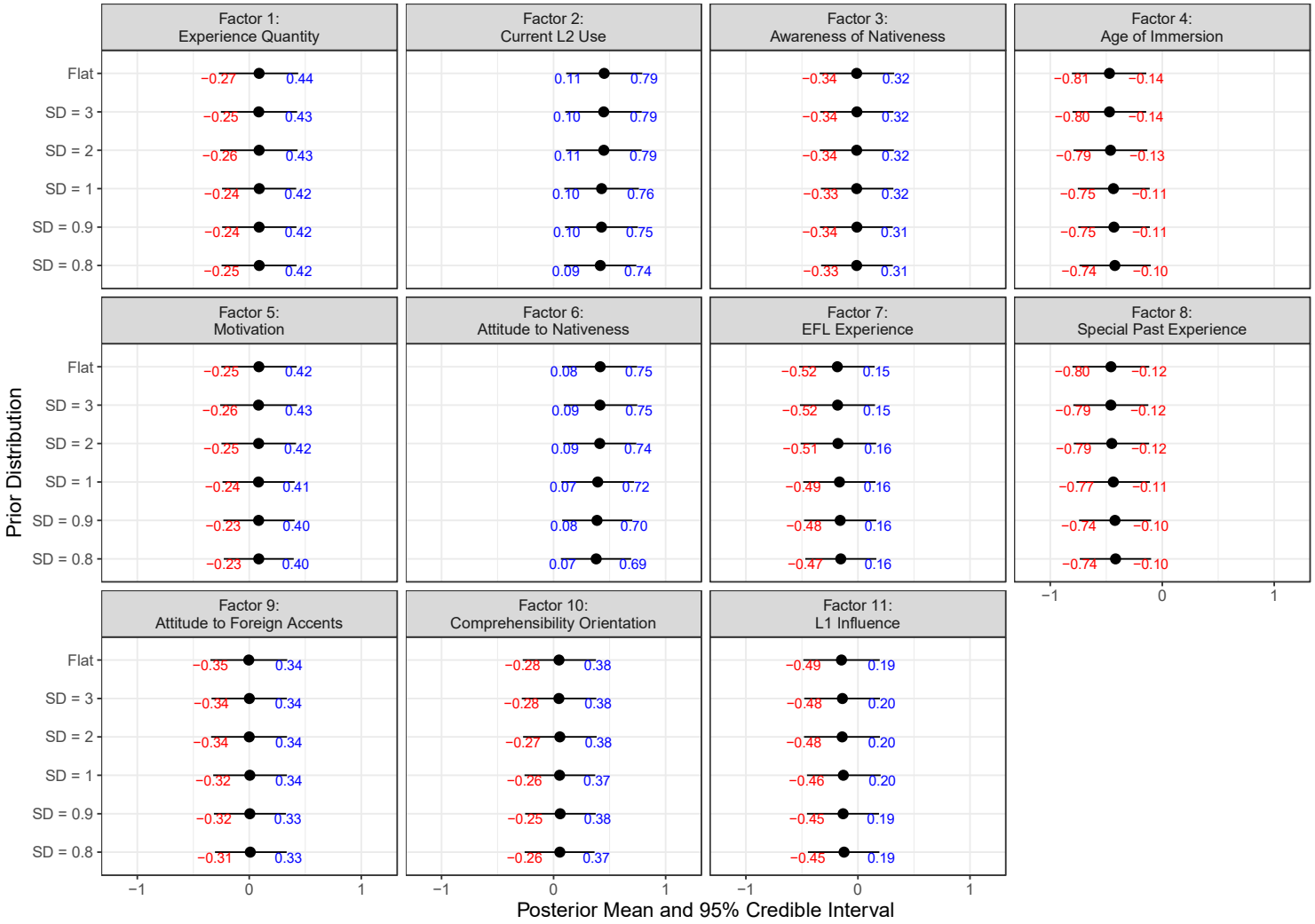

Figure 1. Posterior means and their 95% CIs in each slope parameter across different choices of prior distribution in comprehensibility ratings. SD represents the SD of a normal distribution, while flat corresponds to a flat prior

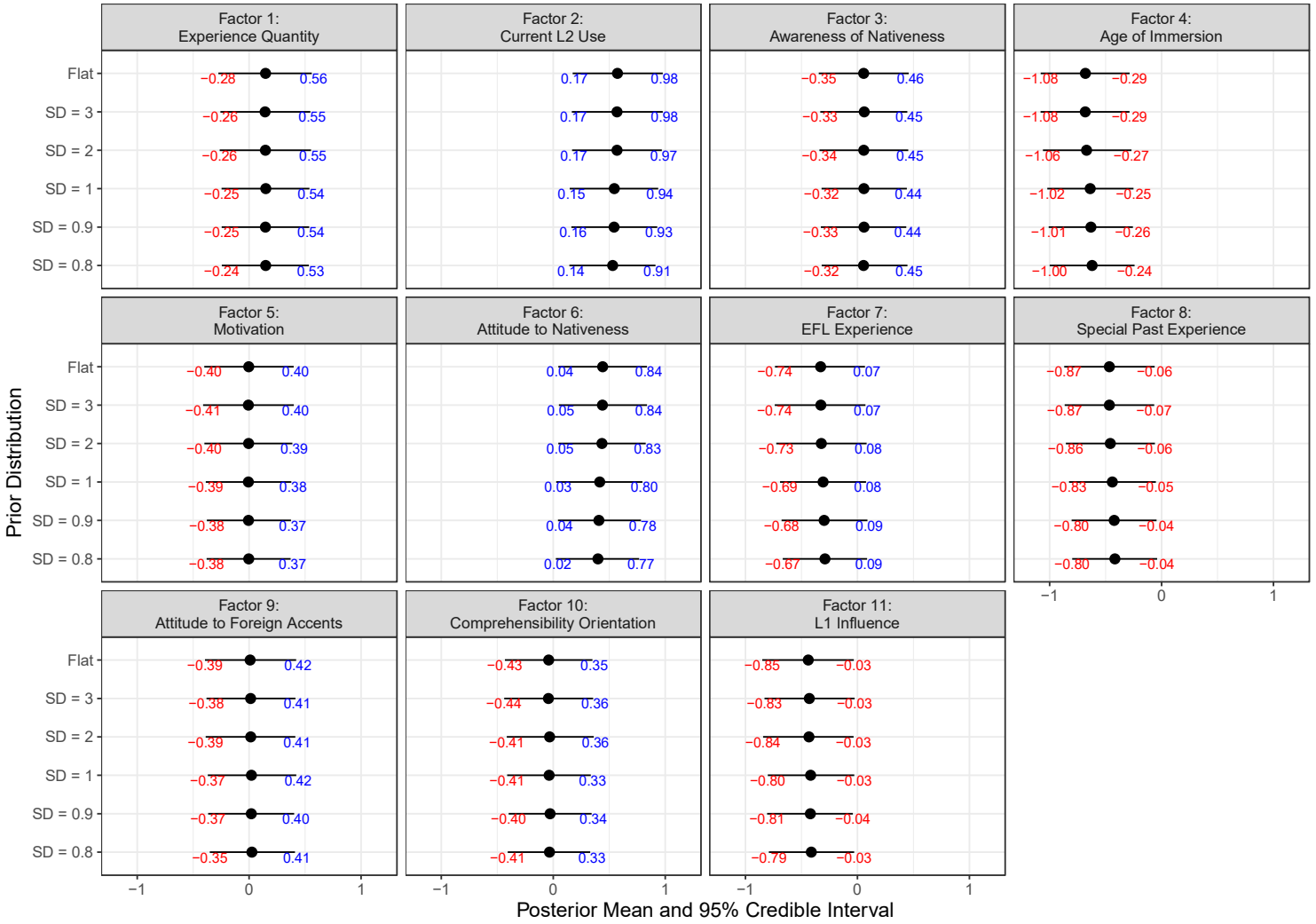

Figure 2. Posterior means and their 95% CIs in each slope parameter across different choices of prior distribution in nativelikeness ratings.

To be published in *Annual Review of Applied Linguistics* (Cambridge University Press)

SUPPORTING INFORMATION-I: Correlates of L2 Comprehensibility vs. Nativelikeness at Different Ability Levels

To grasp the transition of the importance of each factor across the rating scale, for each outcome variable, a series of eight Bayesian mixed-effects binary logistic regression models were fitted to the subset of the data that each only included observations with two adjacent ratings (e.g., 1 vs 2, 2 vs 3, 3 vs 4). The purpose here is to see if and how the posterior mean and its CIs representing the strength of predictors change between low-rating contrasts (e.g., 1 vs 2) and high-rating contrasts (e.g., 8 vs 9). The fixed-effects component of the models included the four (in comprehensibility) or five (in nativelikeness) factors whose 95% CIs did not include 0 in the ordinal regression model, and the random-effects component included by-learner and by-rater random intercepts. The specification of prior distribution is similar to the model we have discussed above, except that flat priors were used for slope parameters.

Figure 1 shows the posterior mean and the 95% CIs of the predictors across the eight models in each outcome measure. The contrast between the two lowest ratings in comprehensibility (i.e., 1 vs 2) is not presented due to the rather small sample size and absurdly large CIs as its results. In the figure, it is, for instance, observed that the effect of Factor 2 (Current L2 Use) is positive in the contrast between 2 and 3 in comprehensibility, which indicates that, in the model that only targeted speech samples rated as 2 or 3, the probability that a sample is given the rating of 3 increases as the factor scores of Factor 2 increase. The effect, however, is trivial, as the 95% credible interval crosses 0. Indeed, most of the models in the figure included 0 in their 95% CIs of the slope parameter, presumably due to the small sample size included in each model. This, however, is not a major issue, as the purpose here is to identify potentially interesting transitional patterns of posterior means.

When each panel of comprehensibility is examined vertically, Factors 2 (Current L2 Use), 4 (Age of Immersion), and 8 (Special Past Experience) do not demonstrate clear patterns, as their posterior means appear to fluctuate randomly from one model to another. In Factor 6 (Attitude to Nativeness), however, there appears to be an increasing trend in that the expected effect of Factor 6 starts at around 0 (i.e., factor scores are unrelated to ratings) and ends positively (i.e., higher scores are associated with higher ratings). Attitude to nativeness, therefore, may influence the discrimination of comprehensibility only at higher ratings. Similarly, regarding nativelikeness ratings, Factors 4 (Age of Immersion) and 11 (L1 Influence) appear to exert stronger influence at higher ratings. The figure, therefore, suggests that the strength of some factors can well vary across the scale of comprehensibility and nativelikeness ratings.

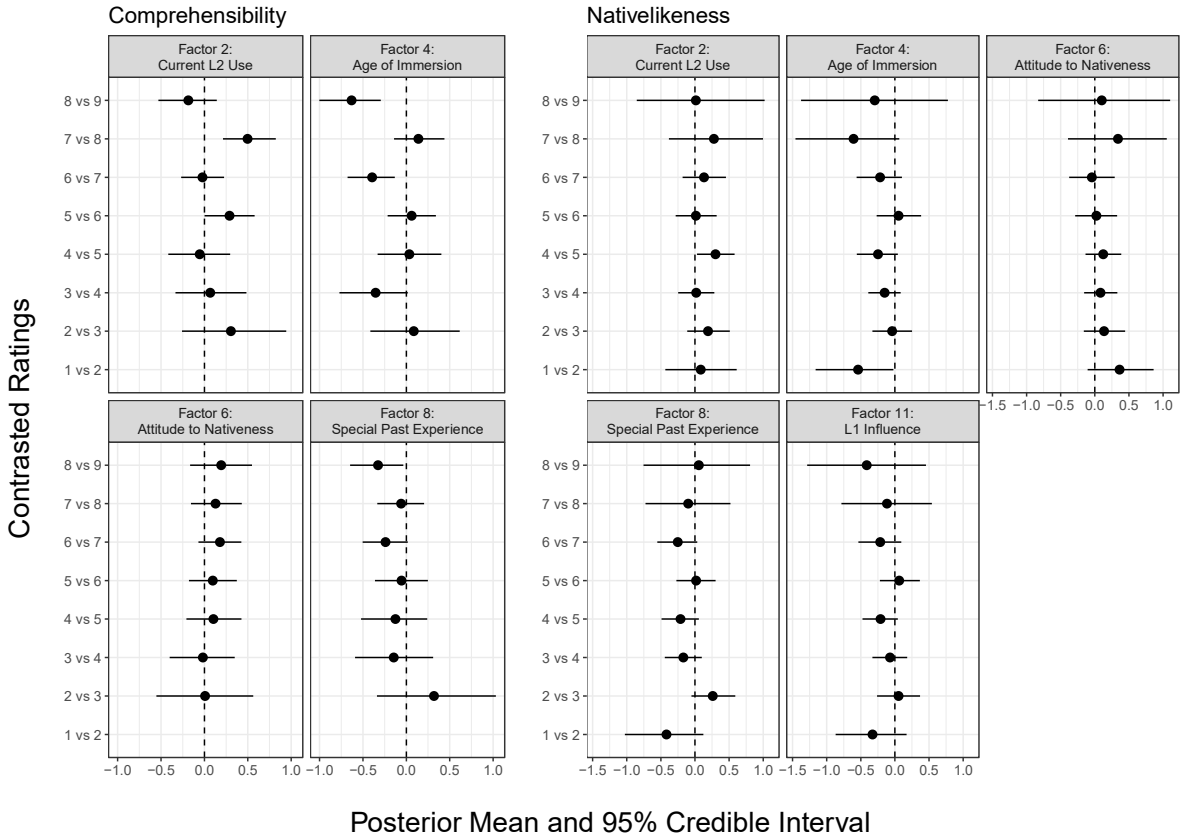

Figure 1. The so-called 'secret weapon' figure (Gelman & Hill, 2007) which shows the posterior means and their 95% CIs in each of the eight Bayesian mixed-effects logistic regression models targeting the observations with two adjacent ratings in each outcome measure.

Reference

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.