Studies in Second Language Acquisition (2025), **47**, 282–307
doi:10.1017/S0272263125000014

RESEARCH ARTICLE

The relationships among L2 fluency, intelligibility, comprehensibility, and accentedness

A meta-analysis

Tuc Chau[1] and Amanda Huensch[2]

[1]English for Multilingual Students Program, University of California Santa Barbara, Santa Barbara, CA USA and [2]Department of Linguistics, University of Pittsburgh, Pittsburgh, PA, USA
Corresponding author: Tuc Chau; Email: tucchau@ucsb.edu

(Received 14 January 2023; Revised 18 December 2024; Accepted 30 December 2024)

Abstract

Fluency, intelligibility, comprehensibility, and accentedness are important dimensions of second language (L2) pronunciation proficiency representing global, listener-based intuitions. This study meta-analyzed 49 reports from 1995 to 2023, examining 141 effect sizes (Pearson *r*) to understand their relationships and possible moderators. Three-level meta-analysis models showed weighted mean correlations of .82, .75, .62, .57, and .32 for fluency/comprehensibility, comprehensibility/accentedness, fluency/accentedness, intelligibility/comprehensibility, and intelligibility/accentedness, respectively. Task types moderated correlations for fluency/accentedness, intelligibility/comprehensibility, and intelligibility/accentedness, with controlled tasks leading to higher correlations. Ratings of multiple dimensions by the same listeners tended to result in weaker correlations for fluency/comprehensibility and comprehensibility/accentedness. The findings imply that having an accent does not mean being unintelligible and support prioritizing intelligible and comprehensible speech over accent reduction. The study also highlights an over-reliance on first language speaker norms in L2 pronunciation research and advocates for more transparent reporting.

Introduction

In second language (L2) pronunciation research, fluency, intelligibility, comprehensibility, and accentedness (FICA) have become increasingly important when it comes to thinking about and measuring speaking performance. One of the seminal studies that pushed forward the field of L2 pronunciation in this direction was Munro and Derwing (1995a), who challenged the assumption that having an L2 accent was equivalent to being unintelligible. Munro and Derwing (1995a) examined the relationships among accentedness, comprehensibility, and intelligibility in the speech of L2 English learners and demonstrated that a strong L2 accent did not necessarily lead to incomprehensible

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

or unintelligible speech despite accentedness negatively correlated with comprehensibility and intelligibility. Such findings fueled a paradigm shift away from L2 pronunciation teaching goals of accent reduction toward prioritizing comprehensibility and intelligibility (Levis, 2020), while triggering a resurgence of interest in L2 pronunciation research in general and FICA in particular. Saito and Plonsky (2019) proposed a framework that conceptualizes these dimensions as global aspects of L2 pronunciation proficiency, observing an increase in the proportion of pronunciation studies adopting FICA dimensions, from 22.7% in 1982–2007 to 32.7% in 2008–2017.

Given the growing body of research involving FICA, there has been interest in bringing together our understanding of these global speech dimensions and how they are related to one another in review chapters (e.g., Munro & Derwing, 2015; Thomson, 2015, 2017). In this work, FICA dimensions are typically identified as related but distinct, for instance: "intelligibility is at least partially independent of many other commonly discussed dimensions of speech, such as accentedness, comprehensibility, fluency" (Munro & Derwing, 2015, p. 379); or similarly, "accent, intelligibility, and comprehensibility [are conceived of] as partially distinct but related dimensions" (Thomson, 2017, p. 13). In related work, some recent meta-analyses have focused on examinations of the specific linguistic attributes that map onto FICA dimensions. For example, Suzuki, Kormos, and Uchihara (2021) examined fluency and explored the relationship between perceived fluency and utterance fluency to better understand which temporal measurements of speech are most related to intuitive listener judgments. In another meta-analysis, Saito (2021) explored which phonological features influenced comprehensibility and accentedness ratings of English as a Second Language speakers and provided compelling evidence that comprehensibility and accentedness are "readily distinguishable" (p. 891) as raters relied on different types of phonological information to make their judgments. Because the scope of previous meta-analyses has been limited to one or two FICA dimensions, it remains unclear to what extent the dimensions are related and whether each dimension should be considered as distinct. To answer these questions, it is necessary to investigate all six possible relationships among FICA dimensions, which is the focus of the current study.

Meta-analyzing FICA relationships contributes to a better theoretical understanding of L2 pronunciation proficiency within the field of L2 acquisition and pronunciation research. Focusing on the strengths of the relationships among all FICA dimensions provides crucial insights: Many have argued that comprehensibility is a useful proxy for intelligibility (e.g., Kennedy & Trofimovich, 2019; Trofimovich, Isaacs, Kennedy, & Tsunemoto, 2022) with accentedness least related to the other dimensions (e.g., Huensch & Nagle, 2021; Munro & Derwing, 2020), but no meta-analysis has yet confirmed this. Lower correlations, such as the weak relationship between intelligibility and accentedness, challenge traditional assumptions that achieving "nativelike" speech is essential for effective communication. Instead, they reinforce the importance of intelligibility and comprehensibility over accent reduction in L2 instruction. In language assessment, knowledge of FICA relationships can guide the development of assessment tools and rubrics that accurately measure pronunciation proficiency. High correlations might suggest the need for simplified, integrated measures, whereas low correlations justify treating dimensions separately to capture distinct aspects of speech proficiency. Furthermore, by considering factors that moderate FICA relationships, such as task type or listener characteristics, assessments can be designed to provide a comprehensive and considerate evaluation of learners' pronunciation skills. Finally, by identifying and categorizing the methodological and reporting practices involved in FICA studies, the current meta-analysis addressed a gap in the existing literature by providing a comprehensive overview of how L2 pronunciation is evaluated across

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

studies. This includes examining reliability reporting, emphasizing the importance of transparency and rigor in research. In sum, the current study synthesized and meta-analyzed previous findings on the strengths of the relationships among FICA and explored a number of moderating variables (e.g., study context, listener background, task type, and assessment setting) to explain potential cross-study variation in the strength of observed correlation coefficients. By gathering empirical evidence from a wide range of studies, the current study enhances our understanding of the complex interplay between FICA dimensions.

Definitions of FICA dimensions

Before the literature on FICA relationships is reviewed, it is necessary to provide brief definitions of each FICA dimension, especially because these terms have been used in multiple ways in the literature. The current study adopts definitions of the global FICA dimensions in line with Saito and Plonsky (2019). To start, fluency has mainly been referred to in two senses: a broad sense and a narrow sense (Lennon, 1990). While in the broad sense, fluency is equated with L2 oral proficiency or even general L2 proficiency, in the narrow sense, it is considered only one component of such proficiency—the fluidity or flow of speech (Housen & Kuiken, 2009; Lennon, 1990; Thomson, 2015). A helpful framework for conceptualizing fluency comes from Segalowitz (2010) and includes cognitive fluency, utterance fluency, and perceived fluency. Cognitive fluency is the efficient mobilization and integration of the cognitive processes responsible for speech production. Utterance fluency comprises temporal, pausing, hesitation, and repair characteristics of speech. Finally, perceived fluency reflects listener judgments of cognitive fluency based on utterance fluency.

Like the other FICA dimensions, perceived fluency represents listeners' experience with an utterance. It also seems to reflect both cognitive and utterance fluency, with empirical evidence showing its ties to utterance fluency (see Suzuki et al., 2021). For these reasons, and in line with the framework proposed by Saito and Plonsky (2019), perceived fluency, as opposed to measures of utterance fluency, was chosen as the fluency dimension of focus in this meta-analysis. Using a listener-based global dimension such as perceived fluency, measured via listeners' scalar ratings, allows for a more logical comparison to accentedness and comprehensibility, which are similarly listener-based global dimensions representing intuitive experiences and reactions to speech. Multiple perceived fluency rating scales can be found in the literature. For instance, Derwing, Rossiter, Munro, and Thomson (2004) had listeners rate speech samples on a discrete scale from 1 (extremely fluent) to 9 (extremely dysfluent), whereas Saito, Trofimovich, and Isaacs (2016) used a 1,000-point sliding scale, arguing that it enables raters to make more fine-grained judgments than Likert scales.

Intelligibility and comprehensibility are two global speech dimensions related to listeners' understanding of speech. Although these two speech dimensions share a focus on listener understanding, they differ conceptually and methodologically in important ways. Intelligibility as defined by Munro and Derwing (1995a, p. 76), refers to "the extent to which a speaker's message is actually understood by a listener", whereas comprehensibility represents "listeners' perceptions of difficulty in understanding particular utterances" (Munro & Derwing, 1995b, p. 291). Researchers have not reached a consensus on how to best measure intelligibility (Kang, Thomson, & Moran, 2018), so its assessment varies from open dictation with word count, cloze tests, focused interviews of listeners to sentence verification and content summaries as well as comprehension questions (Munro & Derwing, 2015). Of these measures, having

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

listeners transcribe speakers' utterances in standard orthography and coding the transcription for exact word match are acknowledged to be the most commonly used (Isaacs & Trofimovich, 2012; Munro & Derwing, 2015).

Different from the assessment of intelligibility, that of comprehensibility concerns scalar ratings (Munro, 2018). The actual scales used have differed: Munro and Derwing used a 9-point scale from 1 (extremely easy to understand) to 9 (impossible to understand), while some studies have opted for different scale labels (e.g., Isaacs & Trofimovich, 2012; Trofimovich & Isaacs, 2012) or lengths (e.g., Isaacs & Thomson, 2013; Saito, Trofimovich, & Isaacs, 2017). In a series of seminal studies, Munro and Derwing provided repeated evidence that intelligibility and comprehensibility were only partially related: Listeners transcribed speech (to reflect intelligibility) and rated how easy it was to understand (to reflect comprehensibility). While accurately transcribed speech was generally rated easy to understand, in some instances, listeners indicated difficulties understanding speech that they could transcribe accurately. Although it might seem logical to focus on intelligibility as opposed to comprehensibility when attempting to measure L2 speech improvements for just the reason that intelligibility reflects actual understanding, Kennedy and Trofimovich (2019) and Trofimovich et al. (2022) put forth compelling arguments for the usefulness of comprehensibility ratings as a measure in L2 speech research: Comprehensibility has the advantage of practicality over intelligibility; it can be judged relatively quickly and reliably using rating scales, while measuring intelligibility requires unique speech content, specific tasks, and time, with scores less reliable across different tasks. Although not identical to intelligibility measures, comprehensibility often shows similar patterns in listeners' understanding and is linked to listeners' processing effort, emotional reactions, and perceptions of speaker credibility, making it valuable for researchers and educators.

The last target dimension in the current meta-analysis was accentedness or the degree of an L2 accent. Accentedness has been defined by different researchers as "the listener's perception of how closely the pronunciation of an L2 speaker mirrors the pronunciation of a native speaker of a given language" (Jułkowska & Cebrian, 2015, p. 141), "how closely the pronunciation of an utterance approaches that of a native speaker" (Kennedy & Trofimovich, 2008, p. 461), and "how strong the talker's foreign accent is perceived to be" (Munro & Derwing, 1995b, p. 291). Like fluency and comprehensibility, accentedness is often measured by listener judgments on rating scales, although the lengths and labels of the scales for accentedness vary from study to study.

Connections between FICA dimensions

Research into FICA has shown that these dimensions are interrelated but partially independent, with correlations varying not only between but also within studies. Munro and Derwing (1995a) laid the groundwork for empirical research into FICA relationships by systematically investigating the relationships among intelligibility, comprehensibility, and accentedness in the speech of L2 learners. They found that although intelligibility and comprehensibility were correlated with accentedness, a strong L2 accent did not necessarily diminish the intelligibility or comprehensibility of L2 speech. Further studies provided additional evidence of the partial independence of these dimensions and examined speakers across proficiency levels and of different L1 backgrounds (Derwing & Munro, 1997; Munro & Derwing, 1995b).

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

As Thomson (2015) noted, there is ample evidence of the connections among intelligibility, comprehensibility, and accentedness, but their association with perceived fluency is not as thoroughly documented. After reviewing the limited evidence available, Thomson (2015) concluded that "fluency is most related to comprehensibility, somewhat related to accentedness, and apparently least related to intelligibility" (p. 217). Nevertheless, such a conclusion has yet to be statistically verified through meta-analysis.

FICA are said to be and commonly accepted as partially independent because the strengths of the correlations vary from rater to rater (Derwing & Munro, 1997; Munro & Derwing, 1995a, 1995b), indicating the contribution of listener factors to the variability of FICA relationships. However, the correlations among FICA dimensions seem to vary not only within but also between studies, to the extent that the direction of the correlation differs. For example, Munro and Derwing (1995b) reported an average correlation strength of $r = .62$ for comprehensibility and accentedness ratings whereas the coefficient was higher in Isbell et al. (2019) at $r = .92$ and has even been reported as negative at $r = -.28$ in Matsuura et al. (2010). Attempting to better understand what might impact such variation across studies, including whether the nature of the relationships vary as a result of methodological differences across studies, is an interesting and useful line of inquiry.

Potential moderators of FICA relationships

Research in L2 pronunciation that reports FICA correlations (henceforth FICA research) involves methodological decisions regarding participants and measurement procedures, which might affect the correlation strengths. The current meta-analysis explored whether methodological differences, categorized as speaker, listener, and measurement variables, have a moderating effect on FICA relationships.

Speaker variables

FICA researchers must first recruit speakers who will produce speech samples for listeners to evaluate. The speakers may be living in a context where the target language is the majority language (L2) or minority language (ML). In the L2 context, enhanced fluency, intelligibility, and comprehensibility might lead to stronger and clearer correlations among these FICA dimensions. This could be because learners in an L2 environment are regularly exposed to the target language in natural, everyday situations (see e.g., Pérez-Vidal & Juan-Garau, 2011), which provides them with consistent and diverse linguistic input. This frequent exposure allows learners to practice and refine their language skills in real time, leading to a more integrated and simultaneous development of fluency, intelligibility, and comprehensibility. Muñoz and Llanes (2014) compared learners in L2 and ML contexts, showing that ML learners tended to retain strong accents due to limited exposure to natural, spontaneous language input. Learners in an ML context may experience more variable development of FICA dimensions due to less frequent and less immersive exposure to the target language. This limited exposure could result in less consistent practice and fewer opportunities to develop these skills in tandem, leading to more nuanced individual differences and less straightforward relationships between the FICA dimensions. Thus, the current study included study context (L2 vs. ML) as a moderator variable.

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

Listener variables

Variables related to the background of listeners have been acknowledged as possible factors influencing bias in language evaluation (e.g., Wheeler & Kang, 2022). For example, the choice of listeners, whether they are L1 speakers of the target language or experts in phonetics/pronunciation, might moderate FICA correlations. L1 listeners may have different perceptual expectations and thresholds for accentedness compared to L2 listeners (see e.g., Kang, 2012). This perceptual variability could moderate the correlation between accentedness and other dimensions (fluency, intelligibility, and comprehensibility). L1 listeners may have a more intuitive grasp of subtle differences in pronunciation and fluency, potentially leading to more consistent ratings across different FICA dimensions. L2 listeners, on the other hand, may focus more on specific aspects of pronunciation that they find challenging or salient, which could influence the strength and nature of the correlations.

Listener experience might also play a significant moderating role in FICA correlations (Isaacs & Thomson, 2013). Experienced listeners, particularly those with expertise in language assessment or phonetics, may exhibit more refined judgments and potentially different patterns of correlation among FICA compared to less experienced listeners. They may also have extensive experience listening to various accents and show higher tolerance for accentedness. Thus, the current study included listener language background (L1 vs. L2) and listener experience (experienced vs. naïve) as moderator variables.

Measurement variables

Apart from participant variables, differences in the measurement of FICA pertaining to speaking task type, task mode, and rating scale might moderate FICA correlations in various ways (e.g., Crowther, Trofimovich, Saito, & Isaacs, 2018; Huensch & Nagle, 2023). Previous studies had speakers perform speaking tasks that can fall into one of the following three categories: controlled, closed, and open tasks (see Suzuki et al., 2021). In controlled speaking tasks, where speakers read predetermined content out loud, correlations among FICA may be more consistent as the vocabulary/syntax of the utterances are provided and do not vary among speakers. Closed tasks with specific prompts may lead to stronger correlations, especially if the constraints of the task standardize language production and evaluation criteria, for instance, by requiring certain vocabulary or structures to successfully complete the task. Open tasks, allowing more freedom in speech, may result in more variable correlations, as speakers' individual styles and linguistic choices can influence how FICA relate. When it comes to task mode (i.e., dialogic vs. monologic), the relationship among FICA might be clearer in monologic tasks due to the lack of interactional dynamics. Dialogic tasks, involving interactions between speakers, may introduce additional variables. Correlations may be influenced by how well speakers negotiate meaning, leading to different patterns in the relationships among FICA dimensions. Thus, the current study included task type (controlled vs. closed vs. open) and task mode (monologic vs. dialogic) as moderator variables.

For fluency, intelligibility, and comprehensibility, researchers usually need to decide the scale that raters will use to evaluate speakers which could impact FICA relationships (Isaacs & Thomson, 2013) as well as whether the assessments will occur in person or online (Nagle & Rehman, 2021). A continuous-point rating scale (e.g., 100- and 1,000-point scales) allows for finer discrimination between levels of proficiency. Strong

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

correlations may emerge as raters can provide more nuanced assessments, capturing subtle variations in fluency, comprehensibility, and accentedness. A discrete-point scale (5-, 7-, or 9-point scale) may lead to clearer correlations if raters align more consistently on specific proficiency levels. However, this may oversimplify the evaluation, potentially resulting in less nuanced correlations. The assessment setting, whether in person or online, introduces different environments that can impact listener judgments and, consequently, FICA correlations. In-person assessments may provide more controlled conditions, allowing for clearer communication and potentially more accurate evaluations of FICA. Online assessments, however, might introduce distractions or technical issues that could affect listener judgments and, consequently, FICA correlations. Thus, the current study included rating scale (discrete-point scale vs. continuous-point scale) and assessment setting (in person vs. online) as moderator variables.

Finally, primary studies have pointed to the potential moderating effect of additional methodological factors, such as including an explanation of FICA to listeners and/or provision of practice items to listeners (Isaacs & Trofimovich, 2012), the number of dimensions investigated (Huensch & Nagle, 2021), the order in which dimensions were rated (O'Brien, 2016), and the inclusion of L1 speaker stimuli (Flege & Fletcher, 1992). Providing clear explanations of FICA dimensions to listeners enhances their understanding and may lead to more consistent assessments. Additionally, offering practice items familiarizes listeners with evaluation criteria, potentially strengthening correlations among FICA dimensions. The number of dimensions investigated and the order in which they are rated might also play a role, affecting the complexity of the assessment task and listeners' attention. For instance, evaluating more dimensions may increase the distinction among dimensions for raters who would be potentially more aware of or focused on attempting to differentiate them. At the same time, rating more dimensions simultaneously might result in a halo effect, or more uniform responses across dimensions and thus higher correlations, if raters do not distinguish well among dimensions or are biased by a general impression of an utterance (see, e.g., Myford & Wolfe, 2003). Consecutive rating may lead to fatigue or habituation, affecting the reliability of judgments and potentially weakening correlations among FICA dimensions. Non-consecutive ratings may mitigate these effects, leading to more consistent assessments. Finally, the inclusion of L1 speaker stimuli provides listeners with a benchmark for comparison and may influence their judgments of L2 speech (Flege & Fletcher, 1992). Exposure to L1 speaker models can affect listeners' perceptions of accentedness and, consequently, the correlations among FICA dimensions. Comparisons between L2 and L1 speech may lead to stronger correlations if L2 speech is consistently evaluated against L1 speaker norms. Thus, the current study included dimension definition (with vs. without definition), practice (with vs. without practice), number of dimensions rated (2 vs. 3 vs. 4), rating order (consecutive vs. non-consecutive), and L1 stimuli (with vs. without L1 stimuli) as moderator variables.

With the growth of FICA studies, we are now at a point where it is possible to bring together this body of research to not only draw conclusions about FICA relationships but also identify their key moderators. In order to do this, three research questions (RQs) were formulated:

RQ1. Which task types, tasks, measures, and instruments are used to assess L2 FICA across pronunciation studies?

RQ2. To what extent are L2 FICA related to each other?

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

RQ3. To what extent do different study contexts, participant backgrounds, and outcome measures moderate the relationships between L2 FICA?

Method

We followed three main steps to answer the RQs: (a) defining a set of inclusion/exclusion criteria and searching for relevant studies; (b) coding the studies for correlation coefficients and features related to study background, participants, and outcome measures; and (c) calculating weighted mean correlation coefficients for the relationships.

Study identification and retrieval

To be included in the meta-analysis, a study had to (a) be a primary study focusing on speech stimuli from adults (18+)[1] who were speaking an L2. Studies that included synthesized speech, used a matched-guise design, or reported on disordered speech were excluded. Additionally, studies had to (b) measure the strength of the relationship between at least two of four FICA dimensions (e.g., fluency and intelligibility, fluency and comprehensibility) and (c) report their correlation coefficients (i.e., Pearson *r*).[2] All the studies were published before April 1, 2023.

Following previous suggestions (Plonsky & Brown, 2015; Plonsky & Oswald, 2012), we tried to be as inclusive as possible in our approach to searching. We first used different combinations of keywords (*fluency, intelligibility, comprehensibility, accentedness, second language/L2*, and *foreign language/FL*) to search library-housed databases, including Academic Search Premier, Education Full Text, Education Resources Information Center, Education Source, MLA International Bibliography, PsycArticles, PsycInfo (via EBSCO—an information service provider), Linguistics and Language Behavior Abstracts, and ProQuest Dissertations and Theses. The same keywords were also used to search the *Journal of Second Language Pronunciation*, the IRIS Repository of Instruments for Research into Second Languages, and Google as well as Google Scholar. We manually searched researchers' websites (Dustin Crowther, Tracey Derwing, Talia Isaacs, John Levis, Murray Munro, Kazuya Saito, Ron Thomson, and Pavel Trofimovich), the International Research Foundation for English Language Education's reference lists (Accentedness, Fluency, Intelligibility, Pronunciation, and Speaking Assessment), and Pronunciation in Second Language Learning and Teaching conference proceedings. Finally, we performed "backward searching" or "reference digging" by examining the references of pronunciation instruction meta-analyses (Lee, Jang & Plonsky, 2015; Saito & Plonsky, 2019).

Our literature search returned 6,297 possible studies for inclusion (which included duplicates across the searches). In a first pass, the authors independently reviewed the title and abstract of each unique study to exclude all obviously unrelated studies. All

[1]Our focus is the dimensions as defined by Munro & Derwing (1995a) and examined in a series of follow-up studies with adult learners. Because of this and a potentially limited sample of non-adult learner studies (see e.g., Lee et al., 2015), we limited our search to only those studies with adult (18+) learner samples.

[2]To provide a more inclusive search, study language was not limited to English. The handful of studies written in a language other than English typically included titles and abstracts in English, and Google Translate was used for those that did not. Only one study (a PhD thesis written in Spanish) was not able to be analyzed.

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

studies identified as potentially relevant by either researcher (n = 527) were then subjected to a second pass in which the full text of each study was reviewed to determine whether it adhered to the inclusion criteria. After this, a total of 103 studies were retained for close coding (see Figure 1 for the study retrieval and screening procedure).

Coding of study features

A detailed coding scheme was developed to further filter studies for the final dataset. Each of the 103 studies was coded in Microsoft Access for five categories of features: (a) bibliographic information, (b) speaker, (c) listener, (d) measurement, and (e) correlation. Table 1 describes the coding scheme in detail. The scheme covered a

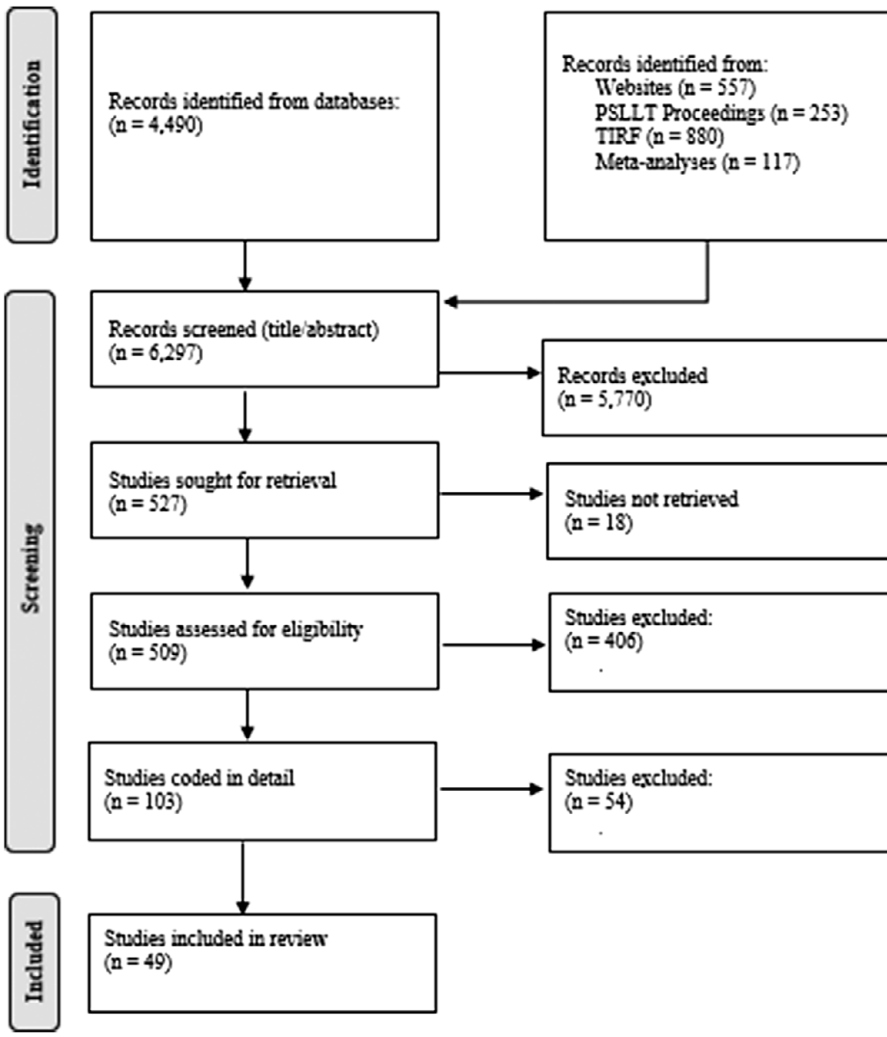


Figure 1. Study retrieval and screening procedure.

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

Table 1. Coding scheme.

Category	Variable	Explanation
Bibliographic Information	Title	Title of the manuscript
	Author(s)	First and last name of all authors
	Year	Year of publication
	Publication type	Conference paper, MA thesis, PhD dissertation, published article
Speaker	Sample size	Number of speakers in the sample
	Age	Mean age or age range of speakers
	L1	First language of speakers
	Target language	Target language of speakers
	Setting	L2, ML, mixed
	Target language proficiency	Beginner, intermediate, advanced, mixed
	Proficiency indicator	CEFR, IELTS, TOEFL, etc.
Listener	Sample size	Number of listeners in the sample
	Age	Mean age or age range of listeners
	L1	First language of listeners
	Language background	Whether listeners are first language speakers of the target language
	Experience	Experienced, naive
Measurement	Task	Read-aloud, picture narrative, monologue, etc.
	Task type	Controlled (e.g., read-aloud, repetition), closed (e.g., picture narrative, TOEFL integrated), open (e.g., interview, monologue)
	Task mode	Monologic (e.g., read-aloud, picture narrative, monologue), dialogic (e.g., conversation, interview, variety show)
	Stimulus length	Mean length of the stimuli in words or seconds
	Number of stimuli	Number of stimuli evaluated by each listener
	Outcome measure	Rating, transcription, etc.
	Instrument	9-point scale, exact match percentage, etc.
	Rating scale	Whether rating scales are discrete-point or continuous-point
	Location	In person or online
	Definition	Whether dimension definitions were provided
	Practice	Whether listeners had practice items
	Number of dimensions	The number of FICA dimensions evaluated
	Rating order	Whether the dimensions were evaluated consecutively
	L1 stimuli	Whether L1 stimuli were included
	Reliability	Inter-rater reliability
Correlation	Correlation type	Pearson, Spearman, etc.
	Correlation coefficient	Value of the correlation

Note: CEFR = Common European Framework of Reference for Languages; IELTS = International English Language Testing System; TOEFL = Test of English as a Foreign Language.

wide range of study characteristics to extract descriptive information and explore potential moderators of the correlations. It underwent several rounds of revision during piloting of a randomly selected sample of studies. As recommended by Plonsky and Oswald (2012), both authors coded the entire sample to ensure inter-rater reliability. Disagreements were discussed and resolved.

During the coding process, additional studies were identified that did not meet the inclusion criteria (because they included non-adult participants, they were duplicate

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

data of another study in the sample, or the correlation coefficient included was not Pearson *r*, etc.). Requests for missing information were sent via email and when necessary, LinkedIn, to 21 authors of whom 14 (66.7%) replied with the requested information.[3] In total, 49 studies together with 141 correlation coefficients (*r*) remained in the final analysis (see https://osf.io/dtp8z/ for a list of included studies).

Analysis

To answer RQ1 (measurement of FICA), we calculated frequencies and percentages of the different task types, tasks, measures, and instruments used to assess FICA. For RQ2 (strengths of FICA relationships), we fitted a three-level random-effects meta-analysis model to each of the six possible relationships (fluency/intelligibility, fluency/comprehensibility, fluency/accentedness, intelligibility/comprehensibility, intelligibility/accentedness, and comprehensibility/accentedness). This approach was taken to account for the dependency between effect sizes occurring when a study contributed more than one effect size to the meta-analysis (Harrer, Cuijpers, Furukawa, & Ebert, 2021). In the current meta-analysis, effect size or correlation coefficient dependence was introduced by researchers who used multiple speaker groups, listener groups, and/or measurement tasks. In the end, five models were fitted using the R package metafor and the function rma.mv (Viechtbauer, 2010). A meta-analysis model could not be fitted for fluency/intelligibility as there was only one study reporting a Pearson correlation (Lee, 2017; $r = .61$, $p < .01$). Correlation coefficients were transformed into Fisher *z* scores for all analyses and converted back for presentation (Borenstein, Hedges, Higgins, & Rothstein, 2021; Cheung, 2015):

$$z = 0.5 \log((1 + r)/(1 - r))$$

The transformed correlations were also weighted by their correspondent sampling variances in order to account for unequal sample sizes:

$$v_z = 1/(n - 3)$$

where *n* is the sample size.

The results were interpreted based on Plonsky and Oswald's (2014) scale for interpreting correlation coefficients in L2 research: small $r \approx .25$, medium $r \approx .40$, and large $r \approx .60$. We addressed RQ3 (moderators of the relationships) by examining I^2, which is the percentage of variability in effects sizes that is due to real differences not sampling error (Harrer et al., 2021). Based on this measure, a predictor can be specified in separate models (i.e., meta-regression) with the *mods* argument of *rma.mv* to help determine which factors are causing between-study heterogeneity. The predictors included in these meta-regression models were study context (L2 vs. ML), listener language background (L1 vs. L2), listener experience (experienced vs. naïve), task type (controlled vs. closed vs. open), task mode (monologic vs. dialogic), rating scale (discrete-point scale vs. continuous-point scale), assessment setting (in-person vs. online), dimension definition (with vs. without definition), practice (with

[3]We would like to thank all of the authors who responded to our queries and provided us with additional information.

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

vs. without practice), number of dimensions rated (2 vs. 3 vs. 4), rating order (consecutive vs. non-consecutive), and L1 stimuli (with vs. without L1 stimuli). Effect sizes were divided into subgroups based on these categories, and a pooled effect size was computed for each subgroup. We included subgroups with as few as two effect sizes (Valentine, Pigott, & Rothstein, 2010).

The data set in this study included both published and unpublished studies in an effort to minimize the potential influence of publication bias that is in favor of positive or statistically significant results (Norris & Ortega, 2006). To test for publication bias, researchers have advocated the use of funnel plots (see Oswald & Plonsky, 2010). Although the use of funnel plots for multilevel data, such as the data in the current analysis, is debated due to the clustering of effect sizes, we included funnel plots to visually assess potential bias (see supplementary material). While clustering may complicate the visualization and interpretation of funnel plots, they still provide valuable information about the distribution of effect sizes. In addition to funnel plots, we extended Egger's test (Egger et al., 1997) by including the standard error (*SE*) of the correlation coefficients as a moderator in the five existing models fitted for RQ2. The intercept of this regression test indicates funnel plot asymmetry or bias: the larger its deviation from zero, the higher the bias, with a significance level set at $p < .10$. We also examined the distribution of correlations for influential outliers. This is defined as Cook's distance greater than 0.45 or hat value greater than $3\frac{1}{k}$, where k is the number of effect sizes (Harrer et al., 2021). All data, code, and output used in this meta-analysis are available at https://osf.io/dtp8z/.

Results

Table 2 provides a summary of the primary studies. The final 49 studies in the current meta-analysis included 50 independent speaker samples and 57 independent listener samples and were conducted between January 1995 and April 2023. A total of 141 correlation coefficients were collected based on these samples. Within individual studies, the speaker sample size ranged from 4 to 120 (M = 29.8; standard deviation [*SD*] = 22.7) with a total of 1,492 speakers included in the current meta-analysis. The listener sample sizes ranged from 2 to 236 (M = 38.8; *SD* = 51.1) with a total of 2,214 listeners. The age of 33 speaker samples (66.0%) for which age information was reported ranged from 18 to 41. Age information was reported for 38 listener samples (66.7%), whose ages ranged from 18 to 75.

In the data set, 33 studies (67.3%) were published articles, 15 (32.6%) were MA and PhD theses and dissertations, and one was a conference paper. The publication language was predominantly English, with the exception of one study published in Korean. English was also the most common target language as reported in 38 studies (77.6%). L2 Arabic, Chinese, Dutch, French, Korean, Russian, and Spanish were the target languages in the remaining 11 studies. Twenty-one studies (42.9%) focused on speakers of the same L1, 27 (55.1%) focused on speakers of different L1s, and one study focused on both single-L1 and mixed-L1 groups of speakers. The speaker L1s in the data set represent a variety of languages: Arabic, Chinese, English, Farsi, Finnish, French, Hindi, Japanese, Kinyarwanda, Polish, Spanish, and Vietnamese. When considering learning context, 26 studies (53.1%) were conducted with speakers in an L2 environment, and 20 (40.8%) were conducted with speakers in an ML environment. One study involved speakers in both L2 and ML contexts. The language acquisition context was not clear in two studies. L1 listeners were recruited as the sole group of listeners in 36

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

Table 2. Summary of primary studies.

Characteristic	Details
Study Period	January 1995-April 2023
Number of Studies	49
Publication Types	Journal articles: 33 MA and PhD theses/dissertations: 15 Conference paper: 1
Publication Languages	English: 48 Korean: 1
Number of Independent Samples	Speaker: 50 Listener: 57
Speaker Sample Size	Range: 4–120 *M*: 29.8 SD: 22.7 Total: 1,492 speakers
Listener Sample Size	Range: 2–236 *M*: 38.8 *SD*: 51.1 Total: 2,214 listeners
Speaker Age	Reported in 33 samples Range: 18–41
Listener Age	Reported in 38 samples Range: 18–75
Target Languages	English: 38 studies Arabic, Chinese, Dutch, French, Korean, Russian, Spanish 11 studies
Speaker L1s	Arabic, Chinese, English, Farsi, Finnish, French, Hindi, Japanese, Kinyarwanda, Polish, Spanish, Vietnamese
Speaker L1 Context	Same L1: 21 studies Different L1s: 27 studies Both single-L1 and mixed-L1 groups: 1 study
Learning Context	L2 environment: 26 studies ML environment: 20 studies Both L2 and ML: 1 study Unclear: 2 studies
Listener Types	L1 listeners: 36 studies L2 listeners: 3 studies Both L1 and L2 listeners: 10 studies
Listener Experience	Reported in 24 studies Naïve listeners: 12 studies Experienced listeners: 9 studies Mixed naïve and experienced listeners: 2 studies Separate groups of naïve and experienced listeners: 1 study
Total Correlation Coefficients	141

studies (73.5%), whereas three studies (6.12%) recruited only L2 listeners. Ten studies (20.4%) recruited both L1 and L2 listeners. Listener experience was reported in 24 studies (49.0%). Of these, 12 studies recruited naïve listeners, nine recruited experienced listeners, two used a mix of naïve and experienced listeners, and one employed separate groups of naïve and experienced listeners.

Measurement of FICA

Among the 49 included studies, 22 (44.9%) measured two FICA dimensions, while 26 (53.1%) measured three, and only one (2.04%) measured four. In 28 studies (57.1%), the dimensions were explained to listeners, and in 27 studies (55.1%), listeners were

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

provided with practice items. The number of practice items ranged from one to 15, with 22 studies including one to five practice items.

A majority of studies (42) took either a simultaneous or a counterbalanced approach to rating the dimensions, whereas the other studies (7) had listeners rate the dimensions consecutively. Regarding speech stimuli, 20 studies (40.8%) included L1 stimuli in the rating/transcription task, while 29 (59.2%) only included L2 stimuli.

The studies included a total of 181 measurements of FICA dimensions. Table 3 provides a detailed description of the task types, tasks, measures, and instruments involved in measuring each of the dimensions. Regarding measurement characteristics, fluency, comprehensibility, and accentedness were measured via rating scales which ranged from 2-, 4-, 5-, 6-, 7-, 9-, and 10-point scales to 100- as well as 1,000-point scales, although the 2-, 4-, 6-, 10-, and 100-point scales were used infrequently (10 times out of 162). The 9-point scale was the most commonly used for assessing fluency, comprehensibility, and accentedness (82 times out of 162) followed by the 1,000-point scale (35 times out of 162). Of the 22 times intelligibility was measured, transcription was used 18 times (81.8%), rating scales were used three times (13.6%), and comprehension questions were used once (4.55%). To elicit L2 learners' speech, researchers employed a variety of tasks. Among them, closed tasks (e.g., picture narratives) were the most commonly used (70 out of 181 times), followed by open (57 times) and controlled tasks (50 times).

Of the 35 times fluency was measured, reliability was reported 26 times (74.3%), and of the 67 times comprehensibility was measured, reliability was reported 48 times (71.6%). The reliability reporting rates of accentedness and intelligibility were lower at 59.6% and 50.0%, respectively. Most studies reported alpha or intraclass correlation (ICC) as their reliability index. The mean (M) reliability coefficients of all FICA measures are shown in Table 4. Finally, 29 studies (59.2%) were administered in person; 13 studies (26.5%) were administered remotely, and 7 studies (14.3%) did not report whether study administration was conducted face to face or online.

Directions and sizes of FICA relationships

Before arriving at the results for RQ2, we tested for publication bias using Egger's tests (see the Analysis section). The results indicated that there was publication bias for the correlations of fluency/comprehensibility ($p < .0001$), intelligibility/comprehensibility ($p = .01$), and comprehensibility/accentedness ($p < .0001$), as demonstrated by the asymmetry in their respective funnel plots (see supplementary material). In contrast, the funnel plots for fluency/accentedness and intelligibility/accentedness displayed more symmetry, and no significant publication bias was detected for these correlations. In addition, we excluded one study from the intelligibility/comprehensibility subset as a result of the outlier analysis[4] (defined as Cook's distance > 0.45 or hat value > $3\frac{1}{k}$). This resulted in increased precision of the average correlation (i.e., narrowed confidence interval [CI]) and reduced variation due to between-study heterogeneity (i.e., change of Q from 52.8 to 17.5, change of I^2 from 81% to 55%). The strength of the relationship also

[4]The study by Jingna and Yao (2013) was excluded due to a high Cook's distance, likely resulting from the exceptionally high correlation between intelligibility and comprehensibility ($r = .95$). This correlation may be due to the study's design, which involved selecting clear, fluent sentences and using a limited number of speech samples, reducing variability in ratings.

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

Table 3. FICA measurement in L2 research.

	Fluency		Intelligibility		Comprehensibility		Accentedness		Total	
	k	%	*k*	%	*k*	%	*k*	%	*k*	%
Task mode										
Monologic	31	**88.6**	18	**81.8**	59	**88.1**	48	**84.2**	156	**86.2**
Dialogic	4	11.4	4	18.2	8	11.9	9	15.8	25	13.8
Task										
Controlled	6	17.1	8	36.4	17	25.4	19	**33.3**	50	27.6
Read-Aloud Paragraph	4	**11.4**			6	8.96	6	10.5	16	8.84
Read-Aloud Sentence			7	**31.8**	8	**11.9**	10	**17.5**	25	**13.8**
Read-Aloud Word			1	4.55	1	1.49	1	1.75	3	1.66
Repetition	2	5.71			2	2.99	2	3.51	6	3.31
Closed	18	**51.4**	4	18.2	29	**43.3**	19	**33.3**	70	**38.7**
Paraphrase			1	4.55	1	1.49	1	1.75	3	1.66
Picture Naming					2	2.99	2	3.51	4	2.21
Picture Narrative	8	22.9	3	**13.6**	16	**23.9**	15	**26.3**	42	**23.2**
TOEFL Integrated	10	**28.6**			10	14.9	1	1.75	21	11.6
Open	11	31.4	10	**45.5**	19	28.4	17	29.8	57	31.5
Conversation	1	2.86	2	9.10	3	4.48	3	5.26	9	4.97
IELTS Long-Turn	1	2.86			1	1.49	1	1.75	3	1.66
Interview	2	5.71	2	9.10	3	4.48	4	7.02	11	6.08
Giving Directions			1	4.55	1	1.49	1	1.75	3	1.66
Monologue	5	**14.3**	5	**22.7**	9	**13.4**	5	**8.77**	23	**12.7**
Presentation	1	2.86			2	2.99	2	3.51	5	2.76
Variety Show	1	2.86					1	1.75	2	1.10
Mixed					2	2.99	2	3.51	4	2.21
SPEAK					2	2.99	2	3.51	4	2.21
Measure										
Comprehension Questions			1	4.55					1	0.55
Rating	35	**100**	3	13.6	67	**100**	57	**100**	162	**89.5**
Transcription			18	**81.8**					18	9.94
Instrument										
2-Point	1	2.78							1	0.54
4-Point							1	1.72	1	0.54
5-Point	3	8.33	1	4.55	8	11.8	4	6.90	16	8.70
6-Point							2	3.45	2	1.09
7-Point	6	16.7	2	9.09	7	10.3	7	12.1	22	12.0
9-Point	12	33.3			35	**51.5**	35	**60.3**	82	**44.6**
10-Point					1	1.47	1	1.72	2	1.09
100-Point					2	2.94	2	3.45	4	2.17
1,000-Point	14	**38.9**			15	22.1	6	10.3	35	19.0
% Accurate			1	4.55					1	0.54
% Exact Match			13	**59.2**					13	7.07
% Keyword			1	4.55					1	0.54
% Nontrivial			4	18.2					4	2.17

Note: The largest percentages of each category are in bold for easy recognition. Issacs and Thomson (2020) used both 5-point and 9-point scales to measure fluency, comprehensibility, and accentedness, resulting in $k = 184$ (rather than 181) for instrument. IELTS = International English Language Testing System; TOEFL = Test of English as a Foreign Language; SPEAK = Speaking Proficiency English Assessment Kit.

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

Table 4. Mean inter-rater reliability coefficients of FICA measures.

			Minimum		Maximum		*M*		*SD*	
	k	Reporting Rate	Alpha	ICC	Alpha	ICC	Alpha	ICC	Alpha	ICC
Fluency	35	74.3%	.81	.90	.99	.98	.93	.95	.05	.02
Intelligibility	22	50.0%	.79	.78	.88	.99	.83	.89	.04	.15
Comprehensibility	67	71.6%	.66	.79	.99	.99	.91	.93	.08	.07
Accentedness	57	59.6%	.56	.80	.99	.99	.91	.92	.11	.08

Table 5. Overall results for FICA relationships.

				95% CI			
	k^a	*r*	SE	Lower	Upper	*Q*	$I^2_{\text{Level 3}}$
Fluency/Comprehensibility	34	.82	.06	.77	.85	62.4*	48
Fluency/Accentedness	19	.62	.08	.51	.71	38.2*	59
Intelligibility/Comprehensibility	17	.57	.09	.42	.69	42.7**	64
Intelligibility/Accentedness	19	.32	.12	.08	.52	40.9*	71
Comprehensibility/Accentedness	50	.75	.08	.67	.81	295.7***	84

Note: The final number of correlation coefficients decreased from 141 to 139 because two studies were excluded from the correlation subsets, with one an outlier and the other the only study investigating fluency/intelligibility.
* $p < .01$;** $p < .001$; *** $p < .0001$.

decreased from $r = .67$ to $r = .59$. Given such changes to the statistical model, the exclusion was applied in all subsequent analyses.

Table 5 summarizes the results regarding the strengths of the correlation coefficients among FICA dimensions in the current sample. The fluency/intelligibility relationship is not included because there was only one correlation coefficient in the sample. The comprehensibility/accentedness relationship had the most frequently reported coefficients ($k = 50$ out of 139) followed by fluency/comprehensibility ($k = 34$). All the FICA dimensions were positively correlated, with correlation coefficients ranging from .32 to .82. While most of the pooled correlation coefficients between FICA dimensions were large (*r* values near and above .60), indicating strong relationships, the correlation between intelligibility and accentedness was small ($r = .32$). Figure 2 illustrates the strengths of the relationships in the form of a mean plot. Of note, the relationship between fluency and comprehensibility ($r = .82$) demonstrated highly consistent findings across studies as indicated by a small ($SE = 0.06$) and comparatively narrow CIs [.77, .85]. Conversely, the relationship between intelligibility and accentedness demonstrated less consistency as indicated by relatively wide CIs [.08, .52]. Forest plots showing the observed effect, CI, and weight of each study for each relationship are available in the supplementary material.

Moderators of FICA relationships

The last RQ, RQ3, examined the variables that may moderate the relationships among FICA dimensions. The I^2 statistic suggests that between-study heterogeneity makes up about 48% to 84% of the total variation in all the relationships (see Table 5), justifying a closer look at the potential contributions of moderators to the relationships.

The moderator analysis revealed that task type significantly influenced the relationships between fluency/accentedness, intelligibility/comprehensibility, and intelligibility/accentedness relationships. For each comparison, the relationships were strongest when

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

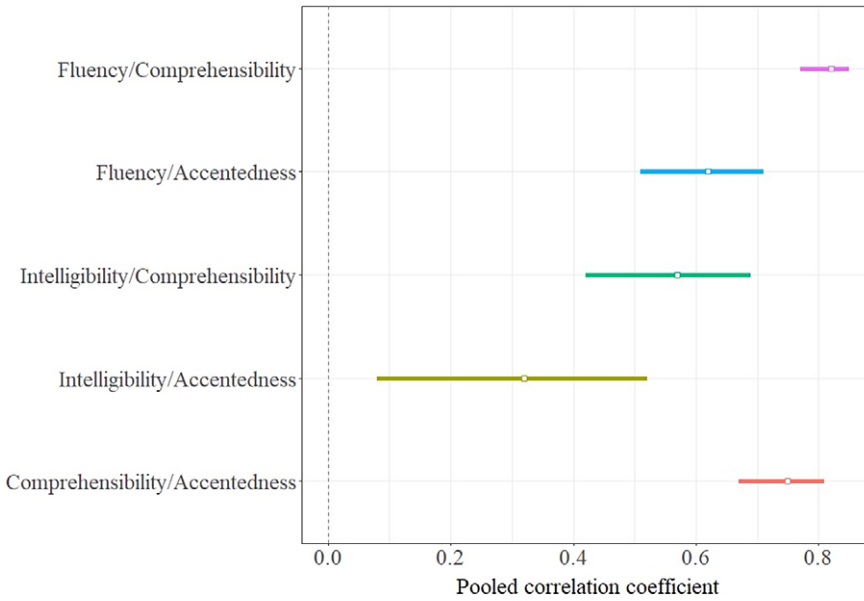


Figure 2. Comparison of FICA relationships.

assessed through controlled tasks (e.g., read-aloud and repetition) and weakest when evaluated with open tasks (e.g., conversation, interview, and monologue). For instance, the fluency/accentedness relationship was stronger in controlled tasks (r = .81, 95% CI [.48, .94], k = 3) versus open ones (r = .52, 95% CI [.34, .67], k = 8), and similarly, for intelligibility/accentedness, there was a medium relationship in controlled tasks (r = .53, 95% CI [.22, .74], k = 7) but a negligible effect in open tasks (r = .04, 95% CI [–.34, .41], k = 8). Additionally, the number of dimensions played a role in shaping the relationships between fluency/comprehensibility and comprehensibility/accentedness. Studies with fewer FICA dimensions measured showed stronger relationships. For instance, the comprehensibility/accentedness relationship was somewhat stronger in studies where they were the only two rated dimensions (r = .84, 95% CI [.74, .91], k = 17) in comparison to those studies when a third dimension was measured (r = .69, 95% CI [.61, .77], k = 32). The same was also true for the fluency/comprehensibility relationship: two dimensions included (r = .88, 95% CI [.80, .93], k = 13) versus three (r = .79, 95% CI [.75, .83], k = 20).

No other significant impacts surfaced related to the other moderator variables. The pooled effects for moderator subgroups in each relationship are presented in the supplementary material. For fluency/accentedness, effect sizes were not pooled in the subgroups of rating order and the number of dimensions measured due to a lack of data. The same holds true for the listener experience subgroups of intelligibility/accentedness.

Discussion

Assessment methods in FICA studies

This study set out to meta-analyze the reported correlations among listener-based global dimensions of L2 speech (i.e., FICA). It first surveyed the methodological and

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

reporting practices of FICA studies. A majority of the studies (74%) employed L1 speakers as the sole evaluators of speech samples. Despite English's status as a lingua franca, leading to its widespread use among L2 speakers, the reliance on L1 speaker norms for measurement within this domain persists (see the *Limitations and recommendations for future research* section). Furthermore, most of the studies (78%) targeted L2 English pronunciation. This reliance may limit the generalizability of FICA findings, underscoring the need for further research in other languages (see also Crowther & Isbell, 2023; Levis, 2021). Such work can identify commonalities and differences in how FICA interact, leading to a deeper understanding of cross-linguistic variation in speech perception and evaluation. Additionally, it can inform language teaching methodologies, curriculum development, and assessment practices tailored to specific linguistic contexts. Fortunately, it appears that the field is moving in this direction: Between 2020 and 2023, five studies focused on a target language other than English, compared to only six such studies observed in the previous period from 1995 to 2019.

A notable finding was that comprehensibility and accentedness were the most frequently measured dimensions ($k = 53$ and $k = 45$, respectively), followed by fluency ($k = 31$) and intelligibility ($k = 17$). This suggests that less attention has been paid to the relationships between intelligibility and other FICA dimensions. This finding might be somewhat surprising due to the increasing acceptance and adoption of intelligibility as a teaching priority (see Levis, 2018). At the same time, intelligibility might be less studied because it is much more difficult to operationalize, both in theoretical and practical terms, in comparison to other FICA dimensions. For instance, although transcription has been used to evaluate this measure of understanding, its coding and analysis are not always agreed upon. Should transcription accuracy scores be based on an exact match of all words or only content words? If a speaker makes a grammatical error and the listener/transcriber "corrects" that error in their transcription, should that count as understanding or not? Relatedly, intelligibility scores often result in very different distributions in comparison to other FICA dimensions (see e.g., Huensch & Nagle, 2021; Munro & Derwing, 1995a), such that they are heavily skewed toward utterances being perfectly intelligible. Ultimately, if most utterances are intelligible, measurements of comprehensibility, which show greater variability and are more reliable across task types (Kang et al. 2018), may be more informative for researchers and teachers attempting to understand and improve L2 pronunciation proficiency (see also Kennedy & Trofimovich, 2019). The variability of intelligibility across tasks and its lack of real-time processing demands in measurement may make it less reliable, especially in capturing the dynamic nature of communication, which might help explain why reliability is reported less frequently in studies on intelligibility.

Another trend observed among the studies pertains to the rating scales utilized. The 9-point Likert scale, used by Munro and Derwing (1995a, 1995b), emerged as the most common scale, often serving as the default measure for fluency, comprehensibility, and accentedness. Although other scales of different lengths were also used, many studies failed to justify their choices. An exception is Saito et al. (2016), who supported their adoption of the 1,000-point slider by arguing that it could facilitate more nuanced judgments. Furthermore, a critical issue in FICA measurement lies in the inconsistent reporting of reliability, with reporting rates varying considerably among FICA dimensions. For example, fluency exhibited the highest reliability reporting rate at 74.3%, while intelligibility had the lowest at 50.0%. Although reliability coefficients reported in FICA studies tend to be higher than those of L2 acquisition instruments (Plonsky &

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

Derrick, 2016), the average reporting rate of 50.2% might undermine the interpretation of FICA results.

Interrelationships among FICA dimensions

The second RQ examined the nature and magnitude of the relationships among FICA. The results confirmed significant associations between these dimensions, with fluency, intelligibility, and comprehensibility inversely related to accentedness. These associations can be attributed to accent features, such as consonant/vowel insertion/deletion, inappropriate syllable reduction, and consonant cluster divergence, which may impede understanding and cause communication breakdowns (Kang, Thomson, & Moran, 2020). However, perhaps of greater interest to pronunciation researchers is the strength of the associations, which can help verify if accentedness "is given more weight than it deserves" (Derwing & Munro, 2009, p. 488). As for this matter, the present study provided empirical evidence on the magnitude of these relationships. Among the aggregated effect sizes, the correlation between intelligibility and accentedness emerged as the weakest, with a correlation coefficient of $r = .32$ (95% CI [.08, .52]). According to Plonsky and Oswald's (2014) field-specific recommendations for interpreting effect sizes, this correlation represents a small effect. This finding so far lends the strongest support to Munro and Derwing's (1995a) preliminary observations that the two dimensions are partially independent.

Unlike intelligibility, fluency and comprehensibility demonstrated strong correlations with accentedness ($r = .62$, 95% CI [.51, .71] and $r = .75$, 95% CI [.67, .81], respectively). One possible explanation for this disparity could lie in the differences in measurement methods between intelligibility and the fluency/comprehensibility/accentedness dimensions. Intelligibility is primarily assessed through transcription tasks, whereas fluency, comprehensibility, and accentedness are measured by rating scales. As pointed out by an anonymous reviewer, these ratings require listeners to make holistic, intuitive judgments and therefore reflect a real-time processing component. As such, the ratings are likely influenced by similar factors (e.g., processing difficulty and alignment with expectations), including linguistic features—an outcome documented in previous work exploring the linguistic features mapping onto these dimensions across tasks (e.g., Crowther et al., 2018). In contrast, intelligibility assessments may not align perfectly with the more fluid and immediate nature of rating fluency, comprehensibility, or accentedness, leading to potential inconsistencies, or incongruencies, in how these components are measured and interpreted within research. In other words, the magnitude of the effect sizes of relationships involving intelligibility might be not only indicative of the theoretical distinctions between the dimensions but also influenced by differences in measurement format. This possibility might lend support to recommendations favoring the use of comprehensibility ratings over intelligibility scores in research and teaching (see e.g., Kennedy & Trofimovich, 2019; Trofimovich et al., 2022).

Another consideration regarding measurement differences between intelligibility and fluency/comprehensibility/accentedness is the potential impact of having listeners transcribe speech. Recent studies have suggested that the inclusion of a transcription task may impact comprehensibility and accentedness ratings along with the strength of their relationship (Huensch & Nagle, 2021, 2023). The logic is that having raters transcribe learner speech might act as an awareness-raising tool—by writing down what they hear, raters are able to differentiate speech that is difficult to understand from

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

that which is merely accented. If we compare comprehensibility/accentedness correlation coefficients in studies with and without a transcription task, we see that studies which included an intelligibility task (k = 15) resulted in weaker comprehensibility/accentedness relationships: r = .62, 95% CIs [0.43, 0.76], compared to studies without an intelligibility task (k = 35): r = .79, 95% CIs [0.71, 0.85]. These findings, combined with the results of the moderator analysis suggesting that including more dimensions weakens the strength of some relationships, call for future work targeting the impact of study design features. Specifically, purposefully examining the impact of including an intelligibility transcription task on comprehensibility/accentedness relationships is an important direction for future research.

While the relationship between intelligibility and comprehensibility yielded a medium-sized correlation coefficient r = .57, the relatively wide CI from r = .42 to r = .69 suggests some uncertainty in the relationship. This may, however, simply mean the number of correlation coefficients examined (k = 17) was insufficient. In fact, intelligibility/comprehensibility had the smallest sample size among the five pairings investigated. Nevertheless, the relationship follows the strength pattern observed in previous research, wherein intelligibility and comprehensibility exhibit a stronger relationship than intelligibility and accentedness (Munro & Derwing, 1995a). The findings underscore the distinction between intelligibility and comprehensibility, despite their intertwined nature. While this relationship does not detract from the benefits of using comprehensibility as an outcome measure for L2 pronunciation research and teaching, it serves as a reminder to researchers in the field that these are indeed distinct dimensions.

Finally, fluency and comprehensibility had the strongest correlation of .82. Moreover, the narrow CI [.77, .85], which is also the narrowest among the relationships, suggests that the correlation was consistently strong across studies. This finding resonates with previous claims that fluency is more strongly associated with comprehensibility than with intelligibility and accentedness (Thomson, 2015). Numerous studies have provided evidence of the close relationship between fluency and comprehensibility. Munro and Derwing (1995a) and Derwing and Munro (1997) reported that goodness of prosody influenced listeners' judgments of comprehensibility. In addition, Derwing, Munro, and Wiebe (1998) and Suzuki and Kormos (2020) found that comprehensibility ratings were linked to factors such as articulation rate and pause duration and ratio. These temporal measures are in turn associated with fluency (e.g., Bosker, Pinget, Quené, Sanders, & de Jong, 2013; Chau, Huensch, Hoang, & Chau, 2022). Moreover, Suzuki and Kormos (2020) showed that not only comprehensibility and fluency were strongly correlated with each other, but also comprehensibility ratings were the strongest predictors of fluency ratings.

Saito and Plonsky (2019) introduced the Framework for L2 Pronunciation Measurement, aiming at assessing the effectiveness of L2 pronunciation instruction. In their framework, FICA are treated as a single measure of "global L2 pronunciation proficiency" on the grounds that it aligns "with empirical evidence in existing L2 pronunciation research" (p. 657). The researchers referred to Derwing et al. (2004) and Munro and Derwing (1995a), who found strong correlations (r >.80) among intelligibility, comprehensibility, and accentedness, as well as between fluency and comprehensibility, and suggested that FICA could be viewed as a "statistically similar phenomenon" (p. 657). The current study uncovered a more nuanced picture of FICA and their relationships. First, intelligibility exhibited the weakest relationship with accentedness. Second, the strength of the relationships can be sorted as follows, starting with the strongest relationship: fluency/comprehensibility, comprehensibility/

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

accentedness, fluency/accentedness, intelligibility/comprehensibility, and intelligibility/accentedness. Finally, fluency and comprehensibility seem to play an intermediary role in connecting the dimensions. They exhibit strong correlations not only with each other but also with intelligibility and accentedness, although information about the fluency/intelligibility pair remains limited.

Factors influencing FICA correlations

Our last RQ was exploratory in nature. It addressed the variability of FICA relationships by examining several potentially influential moderators. These moderators were study context, listener language background, listener experience, task type, task mode, rating scale, assessment setting, dimension definition, practice items, number of dimensions rated, rating order, and inclusion of L1 stimuli. The results revealed significant differences in correlation coefficients based on task type for the correlations between fluency/accentedness, intelligibility/comprehensibility, and intelligibility/accentedness. More specifically, accentedness showed stronger correlations with fluency and intelligibility when measured by controlled tasks (word, sentence, and paragraph reading). An explanation for this finding is the consistency read speech offers, allowing listeners to focus more on the target measures without being distracted by variations in length, content, and grammatical accuracy. These findings echo Crowther et al. (2018), who demonstrated that increased task demands resulted in greater overlap in how linguistic dimensions mapped onto accentedness and comprehensibility, while also leading to higher ratings of both accentedness and comprehensibility. As pointed out by an anonymous reviewer, the linguistic features of the predefined texts in controlled tasks might have downstream impacts on construct validity (if they differentially contribute to the speech dimensions). This provides additional support for the use of open and free tasks in future FICA research.

In addition to the moderating effects of task type, another influential moderator was the number of dimensions measured, which affected the strength of the fluency/comprehensibility and comprehensibility/accentedness relationships. The number of dimensions listeners rate may moderate FICA interconnections because evaluating multiple dimensions simultaneously increases cognitive processing demands, which can lead to variability in judgments as raters attempt to differentiate among linguistic features, particularly when using scales of varying lengths and encountering task-specific influences (Isaacs & Thomson, 2013; Saito et al., 2017; Crowther et al., 2018). The moderator did not significantly impact the fluency/accentedness, intelligibility/comprehensibility, and intelligibility/accentedness relationships, likely due to the limited number of studies ($k = 1$, $k = 2$, and $k = 3$, respectively) that measured only two dimensions in these subsets.

Besides these two moderating effects (which impacted only a subset of relationships), no other significant differences were observed between the subgroups of effect sizes. There are two possible explanations for this result. First, it might be that the relationships among FICA are quite consistent and that the examined variables do not significantly influence the strength of these relationships. However, some studies have indicated influential moderators in specific contexts (e.g., Jułkowska & Cebrian, 2015; Saito et al., 2017; Wheeler & Kang, 2022), as in the case of task type above. Also, certain correlations have relatively wide CIs, suggesting potential variability. Alternatively, it is more probable that between-study differences did not emerge due to the small and

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

unbalanced sample sizes of many moderator subgroups, rendering statistical tests underpowered.

This study advances our understanding of the intricate relationships among FICA, emphasizing their interconnectedness and the factors influencing these dimensions. Notably, intelligibility showed only a weak correlation with accentedness, supporting its partial independence, while fluency and comprehensibility exhibited the strongest interrelation, highlighting their pivotal roles in L2 speech proficiency. The findings of this study both corroborate and challenge existing literature on L2 pronunciation. While they reinforce established relationships, such as the partial independence of intelligibility and accentedness and the strong correlation between fluency and comprehensibility, they also highlight significant gaps and methodological inconsistencies. The underrepresentation of intelligibility, reliance on L1 norms, and inconsistent reliability reporting reveal areas where previous research may have been limited. Additionally, the moderating effects of task type and measurement practices on FICA relationships underscore the need for more rigorous and inclusive methodologies. Together, these findings validate key insights while calling for methodological advancements to better capture the complexities of L2 pronunciation and its dimensions.

Limitations and recommendations for future research

While our study has shed light on the relationships among FICA and their measurement and reporting practices, several limitations warrant acknowledgement. First, the current analysis required studies to report Pearson correlation coefficients, excluding those reporting other types, such as partial and Spearman. This limitation, coupled with potential publication bias evident in three of the five correlation subsets, may weaken result interpretation. Another limitation concerns the robustness of certain moderator analyses, stemming from unequal sample sizes. For instance, the fluency/accentedness relationship subset included 15 L2 studies but only three ML ones (see supplementary material). In addition, several analyses were based on a small number of correlation coefficients. Results from these analyses should be interpreted with caution. Future research with larger sample sizes is needed to further explore the moderators and validate our findings.

Alongside these methodological limitations, our study sample lacked diversity in terms of target languages, with only 10 of 49 studies targeting a language other than English. However, there is a promising trend, with five of 13 studies included since 2020 focusing on a target language other than English, suggesting a shift in research direction. A clear recommendation for future FICA research is the inclusion of participants whose L2 is not English (see also Levis, 2021).

Two design features likely to restrict external validity involve the uses of raters and speaking tasks. In terms of raters, language background (L1 vs. L2) was still the prime parameter for recruitment, despite English being a lingua franca and the diverse interactions English language learners may have with L2 speakers. Having L2 listeners rate speech samples can thus better reflect the diversity of real-life interactions and provide useful insights for the research domain. Regarding speaking tasks, there is a movement towards more authentic tasks beyond controlled read-aloud or repetition tasks, with open tasks favored to measure fluency and intelligibility. Researchers should continue this trend, adopting more spontaneous tasks to enhance ecological validity of findings unless the need for controlled tasks is justified.

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

Lastly, primary studies should demonstrate rigorous and complete reporting practices, including but not limited to reporting reliability estimates. Incomplete reporting may lead to a lack of transparency and reproducibility in research findings, making it difficult for other researchers to assess the reliability and validity of the reported correlations. This could potentially introduce bias or inaccuracies into meta-analytic studies or systematic reviews that rely on the aggregated data. Additionally, the absence of reliability reporting may raise questions about the robustness and credibility of the reported correlations, hindering the identification and correction of methodological flaws or inconsistencies in FICA studies. Promoting comprehensive and transparent reporting practices in FICA research is thus crucial. If space constraints are a concern, researchers can consider the use of online repositories for additional data, such as Open Science Framework. Through these efforts, we can refine our understanding of listener-based global dimensions of L2 speech.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263125000014.

Data availability statement. This manuscript earned the Open Data badge for transparent practices. The data (including the list of included studies) and analysis code are available at https://osf.io/dtp8z/.

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd ed.). Wiley.

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*, 159–175. https://doi.org/10.1177/0265532212455394

Chau, T., Huensch, A., Hoang, Y. K., & Chau, H. T. (2022). The effects of L2 pronunciation instruction on EFL learners' intelligibility and fluency in spontaneous speech. *Teaching English as a Second Language Electronic Journal (TESL-EJ)*, *25*. https://doi.org/10.55593/ej.25100a7

Cheung, M. W.-L. (2015). *Meta-Analysis: A structural equation modeling approach*. Wiley.

Crowther, D., & Isbell, D. R. (2023). Second language speech comprehensibility: A research agenda. *Language Teaching*, 1–17. https://doi.org/10.1017/S026144482300037X

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2018). Linguistic dimensions of second language accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*, *40*, 443–457. https://doi.org/10.1017/S027226311700016X

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*, 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*, 476–490. https://doi.org/10.1017/S026144480800551X

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, *48*, 393–410. https://doi.org/10.1111/0023-8333.00047

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*, 655–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*, 629–634. https://doi.org/10.1136/bmj.315.7109.629

Flege, J. E., & Fletcher, K. L. (1992). Talker and listener effects on degree of perceived foreign accent. *The Journal of the Acoustical Society of America*, *91*, 370–389. https://doi.org/10.1121/1.402780

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). *Doing meta-analysis with R: A hands-on guide*. Chapman & Hall. https://doi.org/10.1201/9781003107347

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*, 461–473. https://doi.org/10.1093/applin/amp048

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

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

Huensch, A., & Nagle, C. (2023). Revisiting the moderating effect of speaker proficiency on the relationships among intelligibility, comprehensibility, and accentedness in L2 Spanish. *Studies in Second Language Acquisition*, *45*, 571–585. https://doi.org/10.1017/S0272263122000213

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*, 135–159. https://doi.org/10.1080/15434303.2013.769545

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

Isbell, D. R., Park, O. S., & Lee, K. (2019). Learning Korean pronunciation: Effects of instruction, proficiency, and L1. *Journal of Second Language Pronunciation*, *5*(1), 13–48. https://doi.org/10.1075/jslp.17010.isb

Isaacs, T., & Thomson, R. I. (2020). Reactions to second language speech: Influences of discrete speech characteristics, rater experience, and speaker first language background. *Journal of Second Language Pronunciation*, *6*(3), 402–429. https://doi.org/10.1075/jslp.20018.isa

Jułkowska, I. A., & Cebrian, J. (2015). Effects of listener factors and stimulus properties on the intelligibility, comprehensibility and accentedness of L2 speech. *Journal of Second Language Pronunciation*, *1*, 211–237. https://doi.org/10.1075/jslp.1.2.04jul

Jingna, L., & Yao, W. (2013). A study of accentedness in the speech of Chinese EFL learners. *Canadian Social Science*, *9*(5), 150–155. http://doi.org/10.3968/j.css.1923669720130905.2799

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*, 249–269. https://doi.org/10.1080/15434303.2011.642631

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, *68*, 115–146. https://doi.org/10.1111/lang.12270

Kang, O., Thomson, R. I., & Moran, M. (2020). Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Applied Linguistics*, *41*, 453–480. https://doi.org/10.1093/applin/amy053

Kennedy, S., & Trofimovich, P. (2019). Comprehensibility: A useful tool to explore listener understanding. *Canadian Modern Language Review*, *75*, 275–284. https://doi.org/10.3138/cmlr.2019-0280

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, *64*, 459–489. https://doi.org/10.3138/cmlr.64.3.459

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*, 345–366. https://doi.org/10.1093/applin/amu040

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*, 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press. https://doi.org/10.1017/9781108241564

Levis, J. M. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation*, *6*, 310–328. https://doi.org/10.1075/jslp.20050.lev

Levis, J. M. (2021). L2 pronunciation research and teaching: The importance of many languages. *Journal of Second Language Pronunciation*, *7*, 141–153. https://doi.org/10.1075/jslp.21037.lev

Lee, H. (2017). An empirical study to rethink the goals and components of teaching Korean language pronunciation. *한국어 교육*, *28*(3), 105–126.

Matsuura, H., Chiba, R., & Matsuda, A. (2010). Evaluative reactions to L2 English: American, Hong Kong Chinese, and Japanese views. *Journal of Commerce, Economics, and Economic History*, *79*(2), 27–38.

Muñoz, C., & Llanes, Á. (2014). Study abroad and changes in degree of foreign accent in children and adults. *Modern Language Journal*, *98*, 432–449. https://doi.org/10.1111/j.1540-4781.2014.12059.x

Munro, M. J. (2018). Dimensions of pronunciation. In O. Kang, R. Thomson, & J. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 413–431). Routledge.

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*, 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*, 289–306. https://doi.org/10.1177/002383099503800305

Munro, M. J., & Derwing, T. M. (2015). Intelligibility in research and practice. In M. Reed & J. M. Levis (Eds.), *The handbook of English pronunciation* (pp. 375–396). Wiley. https://doi.org/10.1002/9781118346952.ch21

Munro, M. J., & Derwing, T. M. (2020). Foreign accent, comprehensibility and intelligibility, redux. *Journal of Second Language Pronunciation*, *6*(3), 283–309. https://doi.org/10.1075/jslp.20038.mun

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386–422.

Nagle, C. L., & Rehman, I. (2021). Doing L2 speech research online: Why and how to collect online ratings data. *Studies in Second Language Acquisition*, *43*, 916–939. https://doi.org/10.1017/S0272263121000292

Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 1–50). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.13

O'Brien, M. G. (2016). Methodological choices in rating speech samples. *Studies in Second Language Acquisition*, *38*(3), 587–605. https://doi.org/10.1017/S0272263115000418

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110. https://doi.org/10.1017/S0267190510000115

Pérez-Vidal, C., & Juan-Garau, M. (2011). The effect of context and input conditions on oral and written development: A study abroad perspective. *IRAL*, *49*, 157–185. https://doi.org/10.1515/iral.2011.008

Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (Or why 18 meta-analyses of feedback have different results). *Second Language Research*, *31*, 267–278. https://doi.org/10.1177/0267658314536436

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, *100*, 538–553. https://doi.org/10.1111/modl.12335

Plonsky, L., & Oswald, F. L. (2012) How to do a meta-analysis. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 275–295). Wiley. https://doi.org/10.1002/9781444347340.ch14

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

Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly*, *55*, 866–900. https://doi.org/10.1002/tesq.3027

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*, 652–708. https://doi.org/10.1111/lang.12345

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*, 217–240. https://doi.org/10.1017/S0142716414000502

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*, 439–462. https://doi.org/10.1093/applin/amv047

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*, 143–167. https://doi.org/10.1017/S0272263119000421

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal*, *105*, 435–463. https://doi.org/10.1111/modl.12706

Thomson, R. I. (2015). Fluency. In M. Reed & J. M. Levis (Eds.), *The handbook of English pronunciation* (pp. 209–226). Wiley. https://doi.org/10.1002/9781118346952.ch12

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press

Thomson, R. I. (2017). Measurement of accentedness, intelligibility, and comprehensibility. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 11–29). Routledge.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*, 905–916. https://doi.org/10.1017/S1366728912000168

Trofimovich, P., Isaacs, T., Kennedy, S., & Tsunemoto, A. (2022). Speech comprehensibility. In T. M. Derwing, M. J. Munro, & R. I. Thomson (Eds.), *The Routledge handbook of second language acquisition and speaking* (pp. 174–187). Routledge.

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, *35*, 215–247. https://doi.org/10.3102/1076998609346961

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Wheeler, H., & Kang, O. (2022). Impact of L2 learners' background factors on the perception of L1 Spanish speech. *Foreign Language Annals*, *55*, 155–174. https://doi.org/10.1111/flan.12588

Cite this article: Chau, T., & Huensch, A. (2025). The relationships among L2 fluency, intelligibility, comprehensibility, and accentedness: A meta-analysis. *Studies in Second Language Acquisition,* 47: 282–307. https://doi.org/10.1017/S0272263125000014

https://doi.org/10.1017/S0272263125000014 Published online by Cambridge University Press