Studies in Second Language Acquisition (2025), 1–34
doi:10.1017/S027226312500018X

METHODS FORUM

Study and instrument quality in perception-based L2 pronunciation research

A methodological synthesis

Maria Kostromitina[1], Ekaterina Sudina[2] and Eman Baghlaf[3]

[1]Duolingo, Inc, Pittsburgh, PA, USA; [2]University of Maryland, College Park, MD, USA and [3]East Carolina University, Greenville, NC, USA
Corresponding author: Ekaterina Sudina; Email: esudina@umd.edu

(Received 09 June 2024; Revised 02 February 2025; Accepted 13 February 2025)

Abstract

This methodological synthesis surveys study and instrument quality in L2 pronunciation research by scrutinizing methodological practices in designing and employing scales and rubrics that measure accentedness, comprehensibility, and intelligibility. A comprehensive coding scheme was developed, and searches were conducted in several databases. A total of 380 articles (409 samples) that employed 576 target instruments and appeared in peer-reviewed journals from 1977 to 2023 were synthesized. Results demonstrated, among other findings, strengths in reporting several listener and speaker characteristics. Areas in need of improvement include (a) more thorough evaluation and reporting of interrater reliability and instrument validity and (b) greater adherence to methodological transparency and open science practices. We conclude by discussing the implications of these findings for researchers and researcher trainers; by raising awareness of methodological and ethical challenges in psychometric research on L2 speech perception; and by providing recommendations for advancing the quality of instruments in this domain.

Keywords: accentedness; comprehensibility; instrument quality; intelligibility; L2 speech; methodological synthesis; reliability; study quality; validity

As part of the methodological reform movement in second language (L2) research (Byrnes, 2013; Gass, Loewen & Plonsky, 2021), there have been calls to examine the quality of studies within and across substantive domains to enhance scientific rigor of research (Plonsky, 2013, 2014, 2024). With increased globalization that enhances L2 communication as well as technological advancements (e.g., automatic speech recognition models), L2 pronunciation is one of the domains that has been gaining steady attention from researchers in various disciplines, including speech and language pathology, language learning and assessment, computational linguistics, sociology, and others. Among the most commonly investigated and influential constructs in pronunciation research have been L2 speech accentedness, comprehensibility, and intelligibility, first introduced by Munro and Derwing (1995). These constructs have been the ever-increasing focus of investigations, particularly since the change of emphasis in L2 teaching from native-like pronunciation to fluency and communicative value of clear, intelligible speech (Levis, 2005, 2020b).

Classified as listener-based constructs, accentedness, comprehensibility, and, in some studies, intelligibility[1] oftentimes rely on scales that measure listeners' perceptions to operationalize them. However, the use of instruments to describe and measure these constructs has been inconsistent (Thomson, 2017), with primary empirical studies employing scales varying in length, item wording, and endpoint descriptors as well as occasionally conflating comprehensibility and intelligibility (e.g., Isaacs & Trofimovich, 2012).

Given the importance of psychometric measurement for the credibility of study findings and conclusions (DeVellis & Thorpe, 2022), this methodological synthesis sets out to examine the quality of instruments used to operationalize accentedness, comprehensibility, and intelligibility in L2 pronunciation research. On a larger scale, the goal of this paper is to examine the overall quality of studies in this domain regarding participant characteristics, instrumentation, and reporting practices, as put forth by Plonsky (2024) in his framework for evaluating study quality in applied linguistics. This framework defines study quality as a combination of methodological rigor, transparency, ethics, and societal value. Methodological rigor in the framework incorporates instrument validity as one of its essential characteristics. The current study makes use of Plonsky's (2024) framework in achieving its goal. In the literature review that follows, we first discuss the main listener perception-based constructs in L2 pronunciation research and the way they are typically measured. We then summarize synthetic and non-synthetic research in study quality before detailing the goal and research questions (RQs) of the present study.

Literature review

Main constructs in L2 pronunciation research

Over the past several decades, pronunciation has been regaining its place in language teaching in part due to the shift from the native (L1)-speaker pronunciation standard to a focus on successful communication and attainment of more achievable goals of comprehensibility and intelligibility (Derwing & Munro, 2005; Derwing, Munro & Thomson 2008; Levis, 2005, 2018, 2020a, 2020b). In other words, rather than emphasizing the importance of sounding like an L1 speaker of a target language, pronunciation instruction has shifted toward the communicative success of an interaction where an L2 speaker's foreign accent does not necessarily inhibit the understanding of their speech. As Saito (2021) note, how this "achievable" goal of L2 pronunciation has been operationalized varies from study to study. Closely related to this movement are three central and partially overlapping constructs that are used to characterize one's L2 pronunciation proficiency—namely, *accentedness*, *comprehensibility*, and *intelligibility* (Derwing & Munro, 1997, 2005; Ghanem, Kang & Kostromitina, 2024). In its original conceptualization by Derwing and Munro (1997; see Thomson, 2017 for a summary), *accentedness* is defined as a property of speech that is perceived as noticeably different from an L1 speaker standard; *comprehensibility* and *intelligibility* are viewed as related

[1] Typically, intelligibility is measured via transcription rather than scalar measures. We discuss this further in the literature review.

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

aspects of pronunciation albeit different in what exactly is measured by each of them. While comprehensibility is described as listeners' perceived effort to understand the speaker's utterance (i.e., how easy or difficult it is for a listener to understand what a speaker is saying), intelligibility, in Munro and Derwing's (1995) view, is the amount of the speaker's production that is understood at the word and utterance level. According to a more recent paper by Munro and Derwing (2015), intelligibility of L2 speech should be a priority in pronunciation instruction, even though comprehensibility is still important for communicative success. Accentedness, in turn, matters insofar as it does not impede intelligibility of speech (e.g., a noticeable foreign accent can be associated with low intelligibility; see Munro, 2008). At the same time, accentedness and intelligibility, while partially overlapping (e.g., word stress or rhythm contribute to both constructs), are also distinct because speech may be accented yet remain highly intelligible and comprehensible (Derwing & Munro, 1997). Shortly after these original definitions had been introduced by Derwing and Munro (1997), other researchers in the domain of L2 pronunciation set forth their own conceptualizations of the three constructs. Thomson (2017) noted that while accentedness had generally been described similarly in the literature, intelligibility and comprehensibility were sometimes conflated (e.g., Isaacs & Trofimovich, 2012; Saito, Trofimovich, Isaacs & Webb, 2016), leading to possible misinterpretation of the results. This confusion is oftentimes related to differences in how the two constructs are measured (e.g., Gooch, Saito & Lyster, 2016; Kennedy & Trofimovich, 2019).

Apart from accentedness, comprehensibility, and intelligibility of L2 speech, pronunciation researchers have also been concerned with fluency. Fluency has typically been operationalized as a temporal measure of speech (e.g., speech rate or mean length of run; see Rossiter, 2009; Saito, Ilkan, Magne, Tran & Suzuki, 2018), even though it is normally presented as one of the core constructs of L2 proficiency, together with complexity and accuracy (Housen, Kuiken & Vedder, 2012). Additionally, as Chau, Huensch, Hoang and Chau (2022) exemplify, fluency can be operationalized as listeners' judgments. This approach reflects the conceptualization of perceived fluency as a sub-construct of pronunciation proficiency (Saito & Plonsky, 2019). Despite the emerging discussion about potentially combining L2 fluency with intelligibility and including this new sub-construct into L2 pronunciation proficiency—together with intelligibility, comprehensibility, and accentedness—the evidence that fluency is indeed related to the aforementioned constructs is scarce (Chau et al., 2022; Derwing & Munro, 1997; Thomson, 2015). Note, however, that L2 comprehensibility has been found to be related to perceived fluency in some studies (e.g., Suzuki & Kormos, 2020). Additionally, a recent meta-analysis by Chau and Huensch (2025) also reported rather strong relationships between fluency and comprehensibility (aggregated $r = .82$) as well as accentedness (aggregated $r = .62$). Nevertheless, in the current study, fluency has not been included as one of the constructs of interest. The next section of the literature review discusses how accentedness, comprehensibility, and intelligibility have been typically measured in L2 pronunciation research.

How constructs are measured in L2 pronunciation research

Although the constructs of accentedness, comprehensibility, and intelligibility are related, there exist different overarching approaches to measuring each of them: perceptual scale-based judgments and transcription. Because accentedness and comprehensibility are perceptual phenomena that work at the utterance level, they are

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

normally evaluated using scales to elicit listeners' intuitive perceptions of one's strength of accent or the amount of effort needed to understand one's speech (Saito, 2021). In most cases, studies have used holistic (impressionistic) instruments to elicit such judgments (e.g., Galante & Thomson, 2016; Huensch & Nagle, 2023). Holistic rubrics usually assign a single score evaluating a certain construct with regard to an entire sound file and are contrasted with analytic rubrics where listeners assign separate scores to different aspects of a construct within a speech excerpt (e.g., Aksakallı & Yağız, 2020; Elborolosy, 2020). As Saito (2021) notes, accentedness has also been referred to as "native-likeness" or "global foreign accent," all reflecting to some degree the extent to which L2 speech is close to phonological norms of L1 speakers (p. 868). Accentedness and comprehensibility judgments, therefore, are thought to represent instant judgments made by interlocutors in real-life oral (intercultural) communication.

In contrast, intelligibility, in line with Munro and Derwing's (1995) approach, while being a perceptual construct that requires both acoustic and cognitive processing, is typically measured via listener transcription. However, some studies (e.g., Christiner, Bernhofs, Sommer-Lolei & Groß, 2023; Kissling, 2014; Meritan, 2022) evaluated intelligibility using scalar ratings to measure listener perceptions of production accuracy. Notably, it has been argued that this approach can introduce bias into intelligibility measurement (Nagle & Baese-Berk, 2022). Intelligibility has been measured in several additional ways (e.g., true/false statements, perception of nonsense sentences, perception of filtered sentences), even though, as Kang, Thomson and Moran (2018) observed, these techniques may not tap into the same construct. Caution must be taken in selecting the technique to ensure that it fits the purpose of a given assessment of intelligibility.

Scale design features

Regarding the length of rating scales to measure accentedness, comprehensibility, and, in some cases, intelligibility, there seems to be great variation in the literature regarding how the instruments have been designed. There are studies that employed 5-point (Kermad, 2021; Isaacs & Thomson, 2013), 7-point (Baills, Zhang, Cheng, Bu & Prieto, 2021; Dai & Roever, 2019; Derwing et al., 2008), 9-point (Munro, 2017), and even 1,000-point scales (Saito, Trofimovich, & Isaacs, 2017; Saito et al., 2019; Tsunemoto, Lindberg, Trofimovich & McDonough, 2022) to register listeners' perceptions of accentedness and comprehensibility. For example, Munro and Derwing (1995) used 9-point semantic differential scales with endpoints labeled as follows: *1 = no foreign accent/extremely easy to understand* and *9 = extremely strong accent/impossible to understand*, for accentedness and comprehensibility. Moreover, differences in endpoint scale descriptors are also quite common (e.g., in Kennedy & Trofimovich, 2008, comprehensibility scale endpoints were *very easy/hard to understand*, while in Saito et al., 2016, the endpoints were *easy to understand/difficult to understand*); these differences, albeit slight, are rarely explained in the literature (Thomson, 2017). The type of scale also seems to matter. Across applied linguistics domains, researchers typically use Likert, Likert-type, semantic differential, and visual analog rating scales. Likert scales, consisting of series of statements with a range of answer options typically from *strongly agree* to *strongly disagree* or vice versa, are common in measuring attitudes, opinions, or behaviors. Likert-type scales are also commonly employed, with the main difference being that their response options are usually worded using descriptors other than *strongly agree/strongly disagree* (e.g., from *not at all* to *very much*).

Semantic differential scales, where bipolar (e.g., *comprehensible* and *incomprehensible; easy to understand* and *difficult to understand*) or unipolar (e.g., *comprehensible* and *not comprehensible; easy to understand* and *not easy to understand*) adjectives are presented at the two ends of the scale, are usually used to measure one's perceptions or reactions. Visual analog, also often referred to as slider scales, are similar to semantic differential ones in their use, but they typically do not have a fixed number of points to choose from; rather, the respondent is encouraged to move the slider to mark their answer on a continuous line. This makes visual analog/slider scales more interactive. Additionally, slider scales can include non-verbal descriptors at each end of the scale (e.g., smiley faces). For a more comprehensive account of various scale types, the readers are referred to DeVellis and Thorpe (2022).

Although scales are arguably the default instruments when it comes to measuring L2 speakers' pronunciation, some studies have employed rubrics instead (e.g., Aksakallı & Yağız, 2020, for comprehensibility; Elborolosy, 2020, for intelligibility). Rubrics are similar to scales in that they must include a measurement scale; however, rubrics are more commonly used in language assessment to evaluate L2 performance or achievement at different proficiency levels; each level of performance is typically accompanied by a detailed description with specific assessment criteria for raters. Rubrics can be analytic, where the measured construct is broken down into subconstructs (e.g., as in Aksakallı & Yağız, 2020, that measured L2 learners' pronunciation based on five criteria, which included "vowels, consonants, intonation, wordstress and comprehensibility," p. 15), or holistic, where a single score is given based on a comprehensive perception of the construct (e.g., Elborolosy's 2020 rubrics assigned separate holistic scores to intelligibility and fluency). For a distinction between holistic versus analytic scoring elsewhere in L2 research (i.e., L2 writing), see Barkaoui (2010).

There has been some examination of the use of scales in L2 pronunciation research. In Yan and Ginther (2017), minimally explicated scales (with only endpoints labeled) have been found to be more attractive for novice raters compared to other types of scales in the study. Several methodologically oriented studies have examined the functionality of rating scales in L2 pronunciation research. Munro's (2017) comparison of different rating scales suggested that a 9-point scale should remain the preferred scale in the domain given its ease of application. However, when Isbell (2017) evaluated a 9-point rating scale for the measurement of comprehensibility and accentedness in Korean learners of English, listeners' post-rating comments uncovered several challenges related to scale length and ambiguity of differences between adjacent points (e.g., "2" and "3"). Analogously, although Isaacs and Thomson (2013) did not detect differences in the mean comprehensibility and accentedness scores between 5-point and 9-point scales, Rasch probability plots suggested that raters in their study were not able to meaningfully differentiate between the points on the longer scale. More recently, Kermad and Bogorevich (2022) found that a shorter 5-point scale is more reliable in measuring speakers' accentedness and comprehensibility.

Listener factors

In addition to scale design choices, studies tend to differ in their composition of listener participant groups in terms of their number, language background, and the amount of training. A major discussion in the domain of L2 pronunciation has concerned the L1 status of listeners and whether listeners' L1 background may affect how they perceive accentedness, intelligibility, and comprehensibility of L2 speech. Some existing studies

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

have collected L1 judgments of L2 speech (e.g., Burda & Hageman, 2005), while others employed either a mix of L1 and L2 listeners (e.g., Chen & Wang, 2016) or a mix of L2 listeners from one or more language backgrounds (e.g., Dai & Roever, 2019). Notably, these methodological choices are largely motivated by study goals. However, as Foote and Trofimovich (2018) note, empirical evidence suggests that (a) L1 speakers generally find L2 speech less intelligible than L1 speech and (b) speakers of a shared L1 may find each other's L2 speech more intelligible (also referred to as the interlanguage speech intelligibility benefit; Bent & Bradlow, 2003); however, this effect may vary depending on listeners' proficiency, context, and other background characteristics. With accentedness and comprehensibility, research on listeners' L1 effect is scarce (e.g., Matsuura, Chiba, Mahoney, & Rilling, 2014; Munro, Derwing, & Morton, 2006; Saito et al., 2019) with findings suggesting significant but modest and non-pervasive effects of L1 background.

Another listener factor in L2 pronunciation studies measuring accentedness, comprehensibility, and intelligibility has to do with rater training (i.e., naïve versus expert listeners). Existing studies have varied in their definition of "listener training." While some have operationalized training as listeners' educational and/or professional background (e.g., being a linguist, having taken a phonology course, having teaching experience), others have provided actual training to listeners in preparation for perceptual ratings (for a discussion, see Kermad, 2021). While several studies have compared the effect of listener training on speech ratings, their findings have been arguably inconsistent. Some reports maintained that trained raters generally showed higher consistency when rating comprehensibility (Saito et al., 2017) and accentedness (Kermad, 2021) as well as producing higher comprehensibility and accentedness ratings overall (Kennedy & Trofimovich, 2008). At the same time, others have found no significant effect of listener training on rater judgments (Isaacs & Thomson, 2013). It is also noteworthy that in L2 pronunciation proficiency contexts, the measurement of the three constructs is different from expert assessments (commonly done in second-language acquisition [SLA] and L2 testing studies) that are done by professional coders using rubrics (e.g., Isaacs, Trofimovich, Yu & Chereau, 2015).

It is evident that scales targeting accentedness, comprehensibility, and, to some extent, intelligibility, have been used somewhat inconsistently in the literature. First, there is variation across studies in how the three constructs are defined, often leading to overlapping interpretations. Within each construct, there exists variation in operationalizations (e.g., using transcription versus scalar judgments to measure intelligibility). Furthermore, there seems to be considerable variation in the scale design, including its length, type, descriptors, and other features at the instrument level as well as methodological choices at the study level, such as listener and speaker characteristics. Moreover, as suggested by existing non-synthetic studies (e.g., Isaacs & Trofimovich, 2016), a pervasive issue is the lack of transparency regarding how definitions and measurements are arrived at, which hinders the comparability and replicability of findings across L2 pronunciation studies.

Existing research on study and scale quality

Study and instrument quality is crucial for the domain of L2 pronunciation just like any other domain in applied linguistics. High-quality studies with well-designed instruments can provide a robust understanding of L2 pronunciation and help gain confidence in the conclusions we draw from research findings. In the context of applied

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

linguistics research, there have been several attempts to define study quality. Plonsky (2013) defined it as adherence to standards of methodological rigor appropriate in given contexts as well as transparent and complete reporting practices. Expanding on this definition, Gass et al. (2021) accounted for aspects of study quality specifically for quantitative research, including estimating effect sizes and prioritizing exhaustive reporting of results as well as sharing study materials. More recently, Plonsky (2024) proposed a framework for study quality that includes four interconnected elements: methodological rigor, transparency, ethics (contrasted with plagiarism, data falsification, and questionable research practices), and societal value. Methodological rigor in this framework involves, among other aspects, instrument quality. Plonsky (2024) notes that in the larger field of applied linguistics, consistent observations have been made about the lack of validity evidence for the instruments, such as scales and proficiency tests (e.g., Ellis, 2021; Norris & Ortega, 2012; Sudina, 2023). Such validity evidence is especially important because its absence casts doubt on the value of study findings.

Although scale and study quality in L2 pronunciation research have not been examined extensively, there is abundant methodological and psychometric literature on scale development and use that SLA researchers can benefit from. DeVellis and Thorpe (2022) highlighted the importance of scale variability, which can be achieved by increasing either the number of items or the number of response options on the rating scale. However, including too many items or response categories "might fatigue or bore the respondents, lowering the reliability of their responses" (p. 107); at the same time, while single-item scales are easier to administer, they may not capture all the details of a given latent construct and can be problematic for reliability calculations. Another issue related to the number of scale response options concerns the extent to which respondents can meaningfully differentiate between them. Although a large number of options may be beneficial for the scale variance, it may result in so-called "false precision" with variability being random rather than systematic (DeVellis & Thorpe, 2022, p. 107). To this end, Menold and Bogner (2016) recommended 5- to 7-point scales that are arguably more optimal for participants' mental effort yet still maintain the necessary psychometric properties. In addition to the number of items and response categories, other features of scale design, such as the presence or absence of a neutral midpoint, response option labeling, and the type of rating scale used (e.g., Likert or Likert-type), tend to have a bearing on scale validity and reliability (Menold & Bogner, 2016). Synthetic research in instrument and study quality, both in the wider domain of applied linguistics as well as in L2 pronunciation, has also revealed several trends that are noteworthy for the current study. Sudina (2021) systematically described and evaluated the state of study and scale quality in L2 anxiety and motivation research and found several limitations, including a narrow range of target languages examined in primary studies (predominantly English) and the lack of diversity in participant population (the majority being college or university students in foreign language contexts). The study also found insufficient reporting of survey design characteristics, issues with missing data handling, and limited instrument availability. Critically, there was a lack of evidence for scale validity (both content and construct) as well as scale reliability, particularly for 1-item scales. Similar findings were observed in Sudina (2023) regarding scale quality in L2 anxiety and willingness to communicate survey research. In the current analysis, we aim to expand Sudina's (2021, 2023) approach to examine study and scale quality in L2 pronunciation research.

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

Validation of instruments has also been central to other areas of applied linguistics such as language testing. Although this area of investigation is outside of the scope of the current study, it is worth mentioning that scholars in this domain have discussed instrument quality in relation to detailed rubrics designed to rate test-takers' performance as well as standardization of ratings by means of using these rubrics. To that end, Knoch, Deygers, and Khamboonruang (2021) conducted a systematic review of studies undertaking scale development specifically for language testing. In that review, the authors provided a framework for scale development and validation, emphasizing the importance of considering test purpose and score use for scale construct validity. Knoch et al.'s (2021) synthesis cited Isaacs and Trofimovich (2012) and Isaacs, Trofimovich and Foote (2018) as exemplary studies in terms of the development and validation of an L2 comprehensibility scale for testing purposes.

Contributing to our understanding of scale and study quality in L2 pronunciation, several meta-analytic and systematic reviews investigated reliability of scales in this domain of research. Saito (2021) emphasized that reliability of L2 speech construct judgments, such as accentedness and comprehensibility, remains an underexplored area. To this end, the author conducted a reliability analysis that showed that comprehensibility and accentedness judgments yield generally high interrater agreement (.90, .91), regardless of listener background (expert versus L2 versus mixed). Lee, Jang and Plonsky (2015), in a meta-analysis of pronunciation instruction research, found an overall mean interrater reliability estimate of .82. This estimate was similar to that found in a meta-analysis of reliability coefficients in L2 research by Plonsky and Derrick (2016). In their study, the median estimate for interrater reliability of L2 pronunciation was .81 (IQR[2] = .20, k = 55). Saito and Plonsky (2019) conducted a reliability generalization meta-analysis (RGM) where they examined variability in interrater reliability estimates in L2 pronunciation teaching research including listeners' impressionistic judgments of global proficiency. Besides finding an overall increase in reliability reporting, the study reported high reliability estimates in the domain (median ranging from .76 to .93), even though there was variability in the estimates depending on the type of production (controlled versus spontaneous) and target of measurement (global constructs versus specific linguistic targets, such as segmental and suprasegmental features). Additionally, the study noted that L2 pronunciation research could apply more sophisticated statistical techniques (i.e., psychometric corrections) when calculating reliability estimates. While these findings seem to offer validity evidence for scales used to measure L2 pronunciation constructs (as reliability is an important aspect of validity evidence; see Chapelle, 2013), a more fine-grained analysis of reliability estimates accounting for design features of scales that measure accentedness, comprehensibility, and intelligibility would provide additional evidence for study and scale quality in this line of research.

The present study

Being at the forefront of linguistic equity and diversity, L2 pronunciation research is committed to promoting successful communication and understanding the role of L2 users' perceived accentedness, comprehensibility, and intelligibility in this process (Levis, 2020a, 2020b). However, little is known about the overall quality of primary studies and instruments used to measure these constructs. Responding to calls by Nagle

[2] Interquartile range.

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

and Baese-Berk (2022), Plonsky (2024), and Sudina (2021, 2023), this methodological synthesis examines study and instrument quality in L2 pronunciation research by scrutinizing methodological practices in recruiting listeners and speakers as well as designing and employing scales and rubrics that measure accentedness, comprehensibility, and intelligibility in an L2 context. The following RQs guided the study:

RQ1: To what extent has the overall scientific rigor of study design been upheld in perception-based L2 pronunciation research with respect to (a) sample characteristics (listeners and speakers) and (b) instrumentation and procedures at the study level?

RQ2: To what extent and by what means has instrument quality been achieved in perception-based L2 pronunciation research regarding (a) instrument design, transparency, and availability; (b) instrument validity; and (c) instrument reliability?

Method

Eligibility

This section outlines the main inclusion and exclusion criteria at both the study and the instrument levels. To address the RQs outlined above, the scope of this methodological synthesis was confined to peer-reviewed articles that focused on accentedness, comprehensibility, and/or intelligibility in an L2 context. While by no means exhaustive, this sample of primary studies is arguably representative of the target domain and allows for accomplishing the main objectives of the synthesis. When determining the scope of the synthesis, we used similar criteria as Plonsky (2013, 2014) and Sudina (2021, 2023). Specifically, to be eligible for inclusion, a study had to (a) represent primary quantitative L2 research into accentedness, comprehensibility, and/or intelligibility (substantive criterion); (b) appear in a peer-reviewed journal (locational criterion); (c) be published in a paginated issue before June 1, 2023[3] (temporal criterion); and (d) include listeners or judges who rated speech samples using scales or rubrics measuring at least one of the target constructs.

Consequently, the following studies were excluded: (a) all syntheses, meta-analyses, non-empirical studies (e.g., theoretical papers, editorials), and primary qualitative articles; (b) book chapters, conference proceedings, M.A. theses, and Ph.D. dissertations; (c) advance online publications that were not yet assigned to an issue by June 1, 2023; (d) articles that did not focus on accentedness, comprehensibility, and/or intelligibility as one of the central constructs and had insufficient information provided for coding (e.g., if an accentedness instrument was only mentioned in the Method section and used as a means to select eligible speakers, and no numerical results were provided; or if a scale measuring comprehensibility was only used in a pilot study to select speakers for the main study); (e) clinical studies with a focus on special populations (e.g., participants with language or cognitive impairments or disabilities); (f) articles that examined comprehensibility in reading or writing rather than speech; (g) articles that had no L2 participants; (h) studies about perceptions of accentedness, comprehensibility, and/or intelligibility in general or self-ratings thereof (without or prior to listening to sound files); (i) studies that did not include target scales or rubrics (e.g., when intelligibility was measured objectively, via a transcription task); (j) studies that employed fully synthesized stimuli rather than those produced by human speakers; and (k) articles that employed speech ratings from another published study.

[3] The final date was determined based on when the searches were conducted.

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

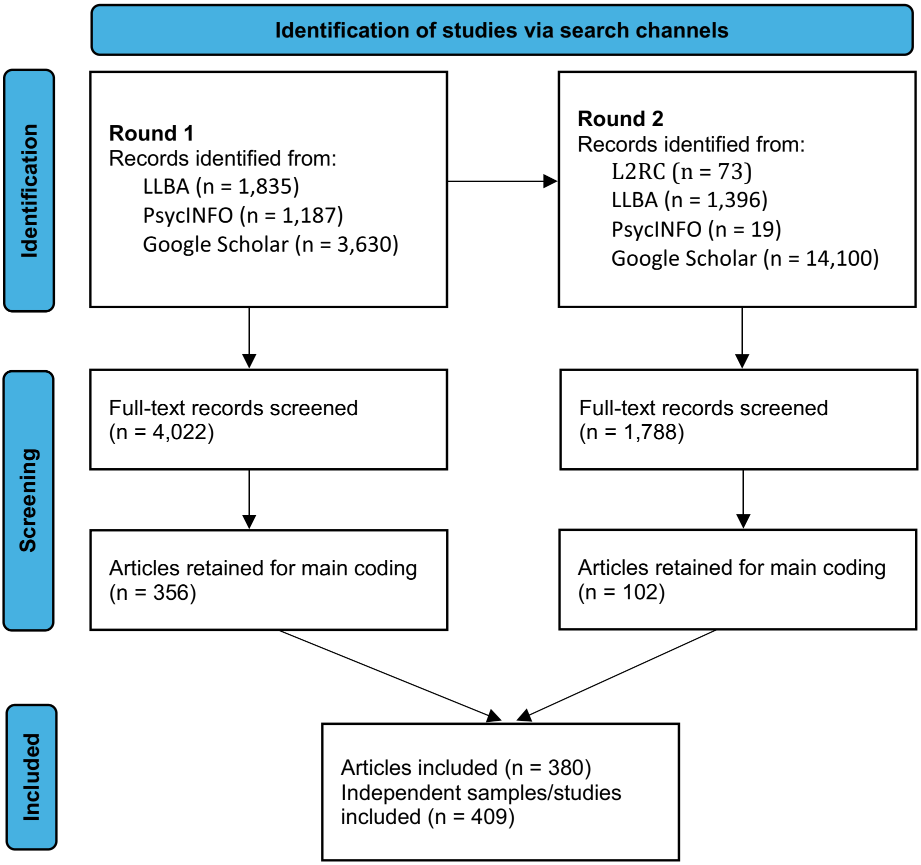
Identification of studies via search channels
Identification
Screening
Included
Round 1
Records identified from:
LLBA (n = 1,835)
PsycINFO (n = 1,187)
Google Scholar (n = 3,630)
Round 2
Records identified from:
L2RC (n = 73)
LLBA (n = 1,396)
PsycINFO (n = 19)
Google Scholar (n = 14,100)
Full-text records screened (n = 4,022)
Full-text records screened (n = 1,788)
Articles retained for main coding (n = 356)
Articles retained for main coding (n = 102)
Articles included (n = 380)
Independent samples/studies included (n = 409)

Figure 1. Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) flow diagram of included and excluded studies. *Source:* Page et al. (2021).
Note: L2RC = Second-language Research Corpus (Plonsky, n.d.).

Searches

This section discusses primary channels and search terms that were used to select eligible articles for this study and provides the results of the search process. As demonstrated in Figure 1, Round 1 included searches in three databases: Linguistics and Language Behavior Abstracts (LLBA), PsycINFO, and Google Scholar (based on recommendations by Plonsky & Oswald, 2015). Round 2 included additional searches in the same databases as well as the Second-language Research Corpus (Plonsky, n.d.). The latter is a comprehensive collection of L2 research articles from 22 journals (*N* = 27,187 files at the time of screening). During Rounds 1 and 2, we screened the full texts of all potentially eligible articles in LLBA and PsycINFO. Due to a large number of hits in Google Scholar, we arranged searches by relevance and conducted full-text screening of the first 100 pages (1,000 articles) in Round 1 and the first 30 pages (300 articles) in Round 2—until there were more false positives than true positives.

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

In Round 1, the following search terms were used: (a) *((accentedness OR comprehensibility OR intelligibility) AND (scale OR questionnaire OR survey) AND (L2 OR "second language" OR second-language OR L3 OR Lx OR LX)) AND stype.exact("Scholarly Journals") AND at.exact("Article") AND la.exact("English") AND PEER(yes)* for LLBA; (b) *(accentedness OR comprehensibility OR intelligibility) AND (scale OR questionnaire OR survey) AND (L2 OR second language OR second-language OR L3 OR Lx OR LX)* for PsycINFO; and (c) *(accentedness OR comprehensibility OR intelligibility) AND (scale OR questionnaire OR survey) AND (L2 OR second language OR second-language)* for Google Scholar. In Round 2, the following searches were conducted: (a) *(accent OR accented NOT accentedness) AND (scale OR rubric) AND (L2 OR L3 OR Lx OR LX) AND stype.exact("Scholarly Journals") AND at.exact("Article") AND la.exact ("English")* for LLBA; (b) *(accent OR accented NOT accentedness) AND (scale OR rubric) AND (L2 OR L3 OR Lx OR LX)* for PsycINFO; and (c) *(accent OR accented -accentedness) AND (scale OR rubric) AND (L2 OR L3 OR Lx OR LX) AND (listeners OR raters OR judges) AND (speakers OR talkers)* for Google Scholar. The L2RC searches were manually performed using AntConc (version 3.5.9) (Anthony, 2020). Combinations of the following search terms were used: *accentedness; heavily accented; comprehensibility; easy to understand;* and *intelligibility.*

After applying inclusion and exclusion criteria during the screening process and removing any duplicates, we identified a total of 458 articles that were eligible for coding. Upon closer examination, the sample was further refined, and the appropriateness of each study was made on a case-by-case basis, which resulted in the exclusion of 78 records (including any duplicates). The final sample comprised 380 peer-reviewed articles (409 independent studies/samples) published between 1977 and 2023 with a total of 576 eligible instruments (scales or rubrics) measuring accentedness (*n* = 310), comprehensibility (*n* = 236), and intelligibility (*n* = 30). Appendix A in Supplementary Materials Online provides a full list of articles included in the synthesis.

Codebook

A codebook was designed to extract the relevant information from primary studies and answer the RQs. It was primarily based on the coding instrument employed in two methodological syntheses of study and scale quality in L2 individual differences (Sudina, 2021, 2023). Other relevant sources in the domain of interest (e.g., Kang & Rubin, 2012; Nagle & Baese-Berk, 2022; Saito, 2021) as well as L2 research more generally (e.g., Plonsky, 2014, 2023; Plonsky & Oswald, 2015) were consulted as well. The codebook consists of Part 1, with Categories 1–4 pertaining to study identification, listeners, speakers, and instruments and procedures at the study level, and Part 2, with Categories 5–8 referring to instrument characteristics, reliability, and content and construct validity. The variables in Part 1 were coded to respond to RQ1, whereas the variables in Part 2 were coded to respond to RQ2. There are 86 variables in total, with both categorical and open-ended items. The instrument was created in Excel and underwent multiple rounds of revisions until the final agreement was reached among the authors (see Appendix B in Supplementary Materials Online; the full codebook has been made available on the IRIS database; https://www.iris-database.org/).

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

Procedure

The screening process began at the end of May 2023 and lasted until September 2023. The coding process was mainly run in parallel and lasted until the end of January 2024. All authors were involved in both the screening and coding procedures. Additionally, a graduate research assistant was trained to assist with one database during the second round of screening. To ensure consistent and accurate coding of the study variables, the authors first coded and discussed two eligible studies together and afterward proceeded to code separately but continued to communicate actively via email as well as during face-to-face and virtual meetings to discuss any pertinent coding issues that arose. Overall, each of the authors coded approximately one third of the sample of primary studies and instruments, after which the lead author additionally double-coded a random subsample of the study (25 studies and 37 scales) to examine intercoder reliability, following Plonsky and Oswald's (2015) recommendations. Intercoder reliability for categorical variables was computed using statistics recommended by Norouzian (2021), such as percent agreement, Cohen's κ, and the S-index using the meta_rate package. For continuously scaled variables, intercoder reliability was assessed using intraclass correlation (two-way mixed, consistency) based on recommendations by Shrout and Fleiss (1979). Intercoder reliability estimates calculated *before* any disagreements had been discussed are demonstrated in Appendix C in Supplementary Materials Online and appear to be satisfactory. (The main points of disagreement were around the type of listeners and rating scales; they were fully recoded in the dataset.) Critically, these estimates were also used for diagnostic purposes, thereby allowing the authors to resolve any discrepancies in the coding sheet and reach the final agreement of 100%. Additionally, several variables in the codebook along with their descriptors were revised and clarified as needed.

Data analysis

To address two RQs that guided the study, various types of descriptive statistics were calculated (e.g., raw frequencies and percentages for categorical variables and means for continuous variables). The purpose of these analyses was to identify and examine existing trends and issues in L2 speech perception research and provide empirically grounded recommendations. For RQ1, the unit of analysis was an independent study (or sample). For RQ2, the unit of analysis was a target instrument. In response to a subquestion on instrument reliability, RGM was conducted (see Plonsky & Derrick, 2016; Plonsky, Marsden, Crowther, Gass & Spinner, 2020). In the first step, reliability coefficients across the scales in the sample were aggregated. A total of eight instruments in the sample had up to four different interrater estimates reported. To avoid redundancy and follow the "one instrument—one reliability estimate" principle when conducting the RGM, the preference was given to those estimates that represented more frequent indices in the sample (e.g., if both Cronbach's α and Spearman's ρ were reported, the preference was given to Cronbach's α). The same approach was used in previous RGMs (e.g., Sudina, 2023). Next, several moderator analyses were conducted to examine reliability estimates by rating scale length, scale type, and scale labeling. To determine which descriptive statistics to report in the RGM (i.e., means and standard deviations (*SDs*) or medians and IQRs), we checked the assumptions of normality and linearity in JASP (JASP Team, 2024) by examining Q-Q plots and running Shapiro-Wilk tests. The data violated the normality assumption; thus, median reliability estimates and their respective IQRs were computed.

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

Table 1. Participant characteristics at the study level (*N* = 409)

Variable	Level	Listeners		Speakers	
		n	%[4]	*n*	%
Type	University students	171	42	144	35
	Unspecified	92	22	106	26
	Mixed	53	13	46	11
	Teachers	40	10	14	3
	General population	17	4	31	8
	Other	16	4	20	5
	Non-university students	15	4	39	10
	Mechanical Turks	5	1		
	Children			9	2
Speaker status	L1 speaker	278	68	26	6
	Non-L1 speaker	65	16	258	63
	Both	54	13	120	29
	Bilingual[a]			5	1
	Unspecified	12	3		
Percentage of participants by gender	Not reported	202	49	127	31
	Reported	207	51	282	69
	Male		37		42
	Female		63		58
	Other		1.6		
Proficiency[b]	Not reported	46	39	173	45
	Reported	73	61	210	55
	Beginner	0	0	15	7
	Intermediate	5	7	39	19
	Advanced	26	36	35	17
	Multiple	42	58	121	58
Age	Mean reported	190	46	205	50
	Mean not reported	219	54	204	50
	SD reported	81	20	89	28
	SD not reported	328	80	320	78
Age by category	Children (age 0–12)	4	1.0	10	2.4
	Teenagers (age 13–17)	3	0.7	18	4.4
	Adults (age 18–54)	185	45.2	200	48.9
	Adults (age 55+)	5	1.2	2	0.5
	Multiple ages	42	10.3	42	10.3
	Not reported	170	41.6	137	33.5
Comparison group	Present	9	2	67	16
	Absent	400	98	342	84

[a] As reported by the primary study authors.
[b] *n* = 119 for the listeners; *n* = 383 for the speakers, excluding L1 speaker groups.

Results

The first RQ inquired about the aspects of study quality at the study level. The sample consisted of a total of 21,866 listeners (21,787 as the main study participants and 79 additional participants across nine comparison groups—i.e., reference groups used as a baseline for analysis, with a mean comparison group size of 8.78) and 19,890 speakers (19,264 as the main study participants and 626 across 67 comparison groups, with a mean comparison group size of 9.48; one group size was not reported). The sample size for listeners as the main study participants ranged from 1 to 1,309 with a median of 24 per study (mode = 10, mean = 53.8, *SD* = 106). The sample size for speakers as the main study participants ranged from 1 to 2,698 with a median of 25 per study (mode = 40, mean = 47.2, *SD* = 163.5).

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

As demonstrated in Table 1, a majority of primary studies in the sample recruited adults in the 18–54 age group (listeners: $M_{\text{age}} = 27.3$; $SD = 5.6$; speakers: $M_{\text{age}} = 26.9$; $SD = 4.1$) and focused on university students (listeners: 42% of a total of 409 independent studies; speakers: 35% of studies). However, general population (e.g., for listeners: Montreal residents; for speakers: actors), mixed groups (e.g., for listeners: English as a second language [ESL] teachers and residents, undergraduate students, and Amazon Mechanical Turk workers; for speakers: high school and university students, L2 English teachers, and students), and other group types (e.g., for listeners: phoneticians, speech-language pathologists, human resources [HR] specialists, and older people residing in assisted-living facilities; for speakers: political leaders, voice actors, management employees, Test of English as a Foreign Language [TOEFL] test-takers, L1 English expatriates living in the Czech Republic, and internationally educated nurses) were also present in the sample. The listeners were predominantly recruited among L1 speakers, whereas the speakers were mainly non-L1 speakers of the target language (in 68% and 63% of independent samples, respectively). Gender was reported in 51% of independent studies for listeners and 69% of studies for speakers; a majority of both listeners and speakers were female. Listeners' proficiency was reported in 61% of independent studies, while speakers' proficiency was reported in 55% of studies; of these, 58% of the sample were at multiple proficiency levels for both listeners and speakers. A comparison group of native listeners was present in approximately 2% of independent studies; a comparison group for the speaker sample was present in 16% of independent studies. The latter comprised predominantly L1 speakers; however, non-L1 and heritage speaker control groups were also present.

As demonstrated in Figures 2 and 3, a majority of studies recruited listeners and speakers from English-speaking countries (USA, Canada, and the UK) and a small fraction of studies recruited participants in more than one country or region; however, participant location was not always available (listeners: 8.3% of studies and speakers: 8.6% of studies).

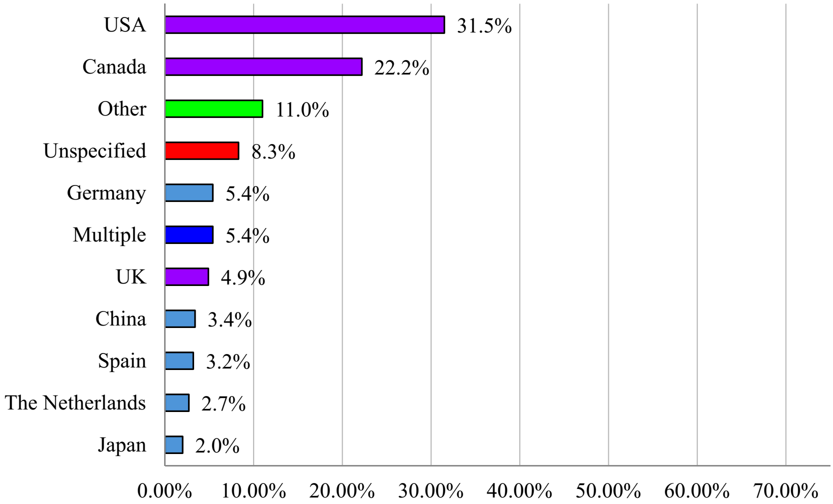

Figure 2. Studies by listeners' location ($N = 409$).
Note: "Multiple" = multi-site locations; "Other" = 30 countries and regions with the lowest frequencies in the sample; "Unspecified" = no listener recruitment location was provided.

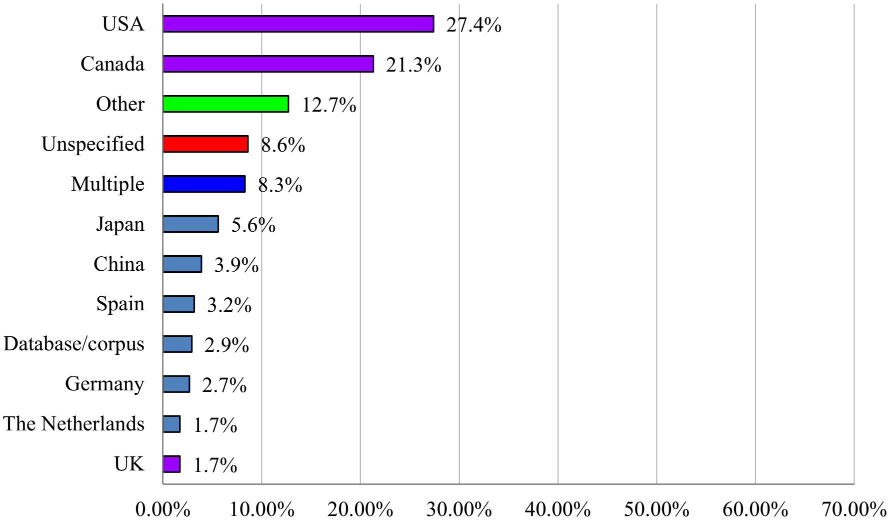

Figure 3. Studies by speakers' location (*N* = 409).
Note: "Database/corpus" = recordings that were taken from an existing database or a corpus dataset; "Multiple" = multi-site locations; "Other" = 32 countries and regions with the lowest frequencies in the sample; "Unspecified" = no speaker recruitment location was provided.

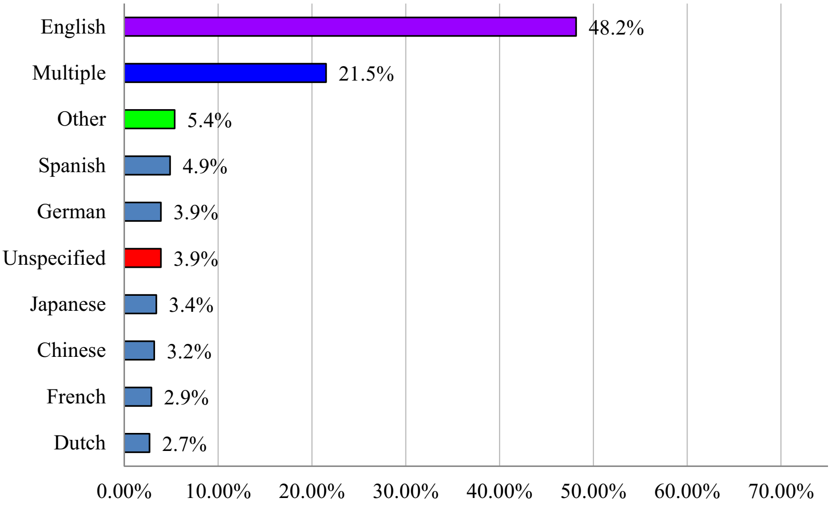

Figure 4. Studies by listeners' L1 language (*N* = 409).
Note: "Multiple" = heterogenous groups of monolingual participants who represented different L1 languages (e.g., English and German); "Other" = < 2% per language in the sample; "Unspecified" = no information about participants' L1 language was available.

As shown in Figure 4, the most common listeners' L1 was English (197 studies), although 88 studies recruited listeners of various L1 backgrounds. Concerning speakers (see Figure 5), only 34 studies recruited L1 English participants; the majority focused on speakers of different L1 backgrounds (223 studies). According to Figures 6 and 7, the

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

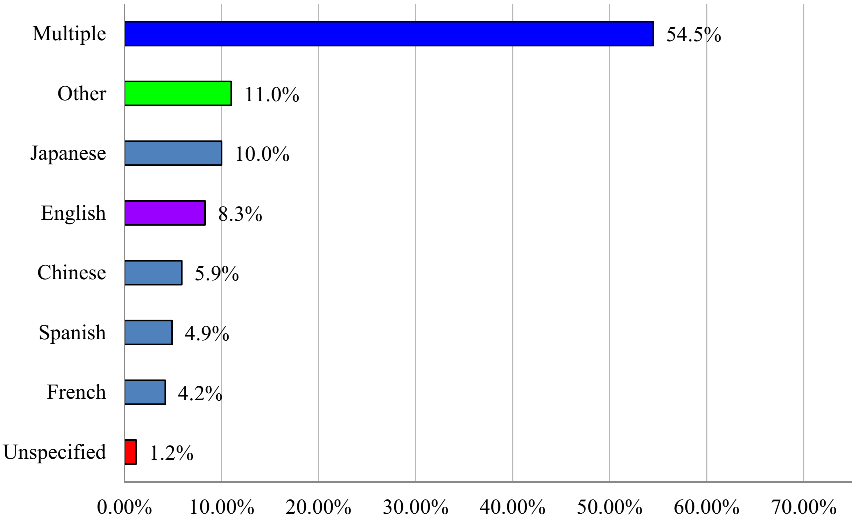

Figure 5. Studies by speakers' L1 language (*N* = 409).
Note: "Multiple" = either heterogenous groups of monolingual participants who represented different L1 languages (e.g., Spanish or Catalan) or bilingual participants with more than one L1 language (e.g., early Spanish-English bilinguals); "Other" = < 2% per language in the sample; "Unspecified" = no information about participants' L1 language was available.

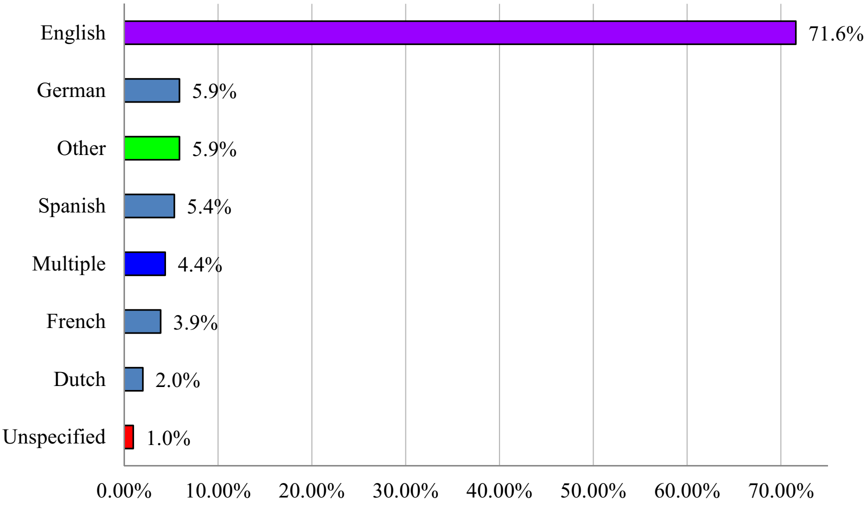

Figure 6. Studies by listeners' target language (*N* = 409).
Note: "Multiple" = more than one target language in a primary study (e.g., Icelandic and English); "Other" = < 2% per language in the sample; "Unspecified" = no information about participants' target language was available.

most frequently investigated target language was English (listeners: 293 studies; speakers: 296 studies), followed by German (listeners: 24 studies; speakers: 24 studies).

Zooming in on the instrumentation, there were 576 eligible instruments in total, including 310 measuring accentedness, 236 focusing on comprehensibility, and

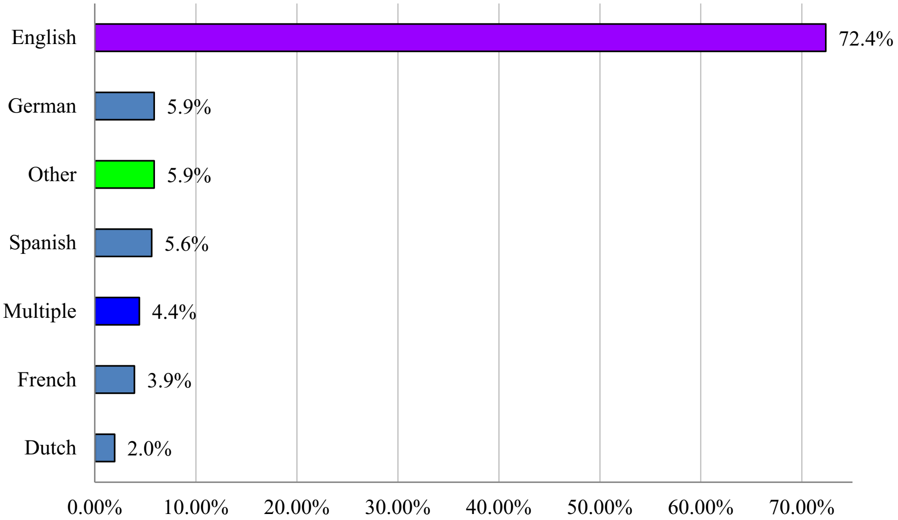

Figure 7. Studies by speakers' target language (N = 409).
Note: "Multiple" = more than one target language in a primary study (e.g., Icelandic and English); "Other" = < 2% per language in the sample.

30 targeting intelligibility. Some studies included more than one target instrument (M = 1.41, SD = .58). Regarding the listening procedure (see Table 2), there was an almost equal number of studies that allowed participants to listen to speech samples more than once and those that did not (approximately 22% in each category). However, more than half of the studies (56%) did not specify this information; in several instances (1%), relistening was not possible (e.g., due to live performance or when raters were asked to evaluate transcribed L2 speech). A small fraction of studies reported whether listeners were allowed to go back and change their rating, but a majority of studies (95%) did not provide this information. Concerning speech file presentation, a majority of studies (69%) reported having (pseudo)-randomized audio samples (or, on several occasions, written transcripts). A handful of studies (2%) explicitly reported not employing the

Table 2. Procedures at the study level (N = 409)

Variable	Level	*n*	%
Relistening allowed	No	88	22
	Yes	90	22
	Not reported	227	56
	Not applicable	4	1
Rating change allowed	No	8	2
	Yes	11	3
	Not reported	390	95
Speech files randomization	No	9	2
	Yes	282	69
	Not reported	111	27
	Not applicable	7	2
Listener training	No	43	11
	Yes	154	38
	Not reported	212	52

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

randomization procedure; yet for another handful of studies (2%), randomization was not possible (e.g., there was only one sound file to listen to or raters were asked to evaluate live performance). Finally, listener training (including rating practice items) was employed in 38% of studies in the sample; no information was available for 52% of independent studies. Note that *listener training* in this study was defined as at least some degree of familiarization with rating materials with or without calibration as opposed to listeners' credentials or professional background; the actual lengths of training varied from a short debrief to a multi-hour training.

The second RQ dealt with speech perception instrument quality features. Table 3 illustrates instrument characteristics pertaining to instrument design, transparency, and availability. The eligible instruments were predominantly scales as opposed to rubrics. Accentedness and comprehensibility instruments constituted a majority of the sample. Several scales that arguably measured intelligibility used the same response options as scales measuring comprehensibility. A majority of instruments targeted independent constructs; however, a handful of instruments (Aksakallı & Yağız, 2020; Baills et al., 2021; Sučková, 2020) focused on accentedness, comprehensibility, and intelligibility as part of a larger construct (e.g., pronunciation, oral performance/proficiency, communicative effectiveness, informational communication). Concerning transparency, a majority of instruments in the sample were not available (approximately 72%); among those that were reported (often without accompanying rating scales), only 17 instruments were made publicly available on IRIS or OSF (Open Science Framework). The number of items was reported for all instruments in the sample. Another transparently reported variable was the number of response options. However, there was a lack of consistency in reporting thereof. For example, several instruments were referred to as 1,000-point scales but ranged from 0 to 1,000, which suggests that they were in fact 1,001-point scales. The least transparently reported variables included pilot-testing procedures, the author(s) and origin of the instruments (newly developed or existing), the type of existing instruments (whether they were modified in some way or not), and the type of adaptations made. Overall, based on the information available, a majority of target constructs were measured using semantic-differential scales, followed by Likert and Likert-type rating scales (for a distinction, see the literature review), and only a few instruments had a neutral midpoint (e.g., *neither agree nor disagree*). A majority of rating scales were accompanied by both verbal and numerical descriptors or only verbal labels on the endpoints. Finally, means and *SDs* of the constructs of interest were reported by the primary study authors more frequently than they were not. Of note, the inclusion of a midpoint has been a controversial issue in the psychometric literature (e.g., Nadler, Weston & Voyles, 2015; Raaijmakers et al., 2000). On the one hand, it may not be ethically appropriate to force respondents to express a definite opinion if they do not feel strongly about a questionnaire item. On the other hand, the midpoint may be taken advantage of by participants who are reluctant to express their judgments, and it can even be understood in different ways (e.g., "neutral" versus "undecided" versus "don't care"; see Nadler et al., 2015; Raaijmakers et al., 2000). As such, scale researchers should weigh the pros and cons of including or excluding the midpoint category and comment on their methodological choices in the research report (DeVellis & Thorpe, 2022).

Moving on to instrument content validity, Table 4 demonstrates that there was insufficient evidence thereof. Specifically, a majority of instruments were single-item scales that arguably cannot adequately represent the complexity and various facets of the constructs of interest. Additionally, only three instruments were examined by experts regarding content validity (i.e., an adapted rubric and two newly developed scales).

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

Table 3. Instrument characteristics (*N* = 576)

Variable	Level	*n*	%
Type	Scale	565	98
	Rubric	11	2
Construct	Accentedness	310	54
	Comprehensibility	236	41
	Intelligibility	30	5
Independent	No, part of a larger construct	20	3
	Yes	556	97
Availability	No	414	71.9
	Yes, in the article/online suppl.	145	25.2
	Yes, on IRIS/OSF	2	0.3
	Yes, in the article/online supplement and on IRIS/OSF	15	2.6
Number of items	Reported: *M* (*SD*) = 1.16 (.86) Range = 1–10	576	100
Pilot-testing	Yes	35	6.1
	Not reported	541	93.9
Author	Reported	123	21
	Not reported	453	79
Origin	New	5	1
	Existing	118	20
	Not reported	453	79
Existing instrument type[a]	Borrowed	13	11
	Adapted (modified)	45	38
	Not reported	60	51
Adaptations[b]	Specified	6	13
	Not specified	39	87
Adaptation reporting[c]	Changed the number of response options	3	50
	Changed the rating scale	3	50
Number of response options	Reported (mean = 89.9, median/mode = 9, range = 2–1,001)	572	99
	Not reported	4	1
Response format	Semantic differential	272	47
	Likert/Likert-type	187	32
	Slider	79	14
	Percentage scale	3	1
	Binary	4	1
	Some rubric[d]	4	1
	Not reported	27	5
Response option labeling	Partially verbal and numerical	355	62
	Partially verbal (endpoints)	119	21
	Fully verbal	28	5
	Fully verbal and numerical	27	5
	Partially verbal and emojis	20	3
	Numerical	7	1
	Emojis	6	1
	Not reported	14	2
Neutral midpoint	Yes	11	2
	No	333	58
	Not reported	153	27
	Not applicable	79	14
Mean	Reported	424	74
	Not reported	152	26
SD	Reported	317	55
	Not reported	259	45

[a] Based on a total of 118 existing instruments;
[b] based on a total of 45 adapted instruments;
[c] based on a total of 6 adaptations reported;
[d] these rubrics were not accompanied by traditional rating scales and were therefore placed in a separate category.

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

Table 4. Instrument content validity (*N* = 576)

Variable	Level	*n*	%
Single-item instrument	No	27	5
	Yes	549	95
Item evaluation	Expert review	3	.5
	Not reported	573	99.5

Table 5. Instrument construct validity (*N* = 576)

Variable	Level	*n*	%
Rasch analysis	Yes	9	2
	No	567	98
Factor analysis results	Some factor analysis	1	.2
	PCA	1	.2
	EFA	2	.3
	No	572	99.3
Factor analysis justification[a]	Yes	4	100
	No	0	0
Measurement invariance	Yes	0	0
	No	576	100
Convergent validity	Yes	0	0
	No	576	100
Divergent/discriminant validity	Yes	0	0
	No	576	100
Validity reference	Reported	13	2
	Not reported	558	97
	Not applicable (new)	5	1

Note: [a]Based on a total of 4 factor analysis results reported.

Content validity of three other new scales and the remaining adapted instruments was not reported to have been evaluated.

As shown in Table 5, there was limited evidence of instruments' construct validity. No instrument in the sample was tested for measurement invariance or examined for convergent or discriminant validity. Few instruments were assessed using Rasch analysis or factor analysis. However, when factor analysis was performed, the choice of a specific technique (e.g., exploratory factor analysis [EFA] or principal components analysis [PCA]) was usually justified. Finally, several existing instruments, although not scrutinized for sample-specific construct validity by the primary study authors, were accompanied by a validity reference from a previous study that served as indirect validity evidence (e.g., "This scoring rubric has been validated in a study by Riaz, Sham and Riaz"; Sheredekina, Karpovich, Voronova & Krepkaia, 2022, p. 6).

Along with scale validity, the second RQ examined the status quo of instrument reliability. According to Table 6, item-total correlations were reported, either fully or partially, for only 2 of 27 multi-item instruments in the sample (see the full list in Appendix D in Supplementary Materials Online). Approximately 52% of the instruments were presented with interrater reliability; Cronbach's α followed by intraclass correlation coefficient were the most reported indices. A total of 18 instruments had other types of reliability reported in addition to or instead of interrater reliability (e.g., internal consistency, test-retest, intrarater reliability). Cronbach's α and intraclass correlation coefficient were, again, the most frequently reported indices in this category. All but five instruments in

Table 6. Instrument reliability (*N* = 576)

Variable	Level	*n*	%
Item-total correlation[a]	Full (corrected)	1	4
	Partial (mean)	1	4
	Not reported	25	93
Interrater reliability	Yes	299	52
	No	277	48
Interrater reliability index[b]	Cronbach's α	150	48.2
	Intraclass correlation	99	31.8
	Unspecified	24	7.7
	Pearson's correlation	17	5.5
	κ	10	3.2
	Interclass correlation	5	1.6
	Spearman's ρ	3	1.0
	Percent agreement	2	.6
	Krippendorff's α	1	.3
Reliability other	Yes	18	3
	No	558	97
Other indices[c]	Cronbach's α (internal consistency)	11	57.9
	Intraclass (intrarater)	3	15.8
	Rasch	2	10.5
	Unspecified (internal consistency)	2	10.5
	Test-retest	1	5.3
Number of subscales	Unidimensional	571	99
	Multidimensional	5	1
Reliability subscales[d]	Yes	5	100
	No	0	0

[a]Based on a total of 27 scales with > 1 item;
[b]based on a total of 311 indices for 299 scales with reported reliability—some instruments had > 1 index reported;
[c]based on a total of 19 alternative reliability indices reported for 18 different scales: 1 scale had 2 such indices reported;
[d]based on a total of 5 multidimensional scales.

the sample were unidimensional. Subscale reliability was reported for all five multidimensional scales.

In addition to examining instrument reliability reporting practices, we conducted a small-scale RGM (see McKay et al., 2021; Plonsky et al., 2020; Sudina, 2021, 2023) to examine whether certain design features of scales used to measure the constructs of interest affect their reliability. Specifically, we compared two most commonly employed interrater reliability indices—Cronbach's α and intraclass correlations—moderated by rating scale length (i.e., 5, 7, 9, 100, and 1,000[5] points), scale type (i.e., semantic differential, Likert/Likert-type, and visual analog), and scale labeling (i.e., fully verbal, fully verbal and numerical, partially verbal and numerical, and endpoints only). Because of how reliability was reported in our sample of primary studies, only the most representative moderators were included.

Table 7 reports the results of the RGM overall and as differentiated by moderator variables. Because sample sizes for each feature varied and were sometimes quite small, the results presented below need to be taken with caution. As seen in the table, there was not much variability in reliability estimates by index type or selected scale design

[5]While these instruments were referred to as 1,000-point scales by primary study authors, many of these scales were, in fact, 1,001-point scales, where 0 appeared on the left side of the scale and 1,000 on the right side of the scale.

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

Table 7. Reliability generalization meta-analysis results

	Cronbach's α	ICC
Moderators	*Mdn, IQR* (*n*)	*Mdn, IQR* (*n*)
Construct		
Accentedness	.93, .08 (71)	.94, .08 (53)
Comprehensibility	.92, .07 (75)	.94, .14 (39)
Scale length		
5-point	.85, .2 (11)	.94, .38 (13)
7-point	.88, .13 (20)	.93, .07 (11)
9-point	.94, .06 (64)	.93, .13 (41)
100-point	.85, .11 (4)	.97, .03 (10)
1,000-point	.93, .05 (34)	.91, .11 (7)
Scale type		
Semantic differential	.92, .08 (68)	.93, .15 (52)
Likert/Likert-type	.92, .09 (41)	.94, .11 (22)
Visual analog	.93, .05 (33)	.97, .07 (15)
Scale labeling		
Fully verbal	.87, .09 (5)	.83, .36 (5)
Only endpoints	.90, .12 (24)	.96, .06 (30)
Fully verbal and numerical	.94, .23 (3)	NA
Partially verbal and numerical	.93, .08 (92)	.94, .10 (55)

Note: ICC = intraclass correlation; IQR = interquartile range; *Mdn* = median; *n* = number of scales; NA = not available.

features. Only the preference for scale labeling somewhat affected interrater reliability, with both Cronbach's α and intraclass correlation being the lowest when fully verbal labeling was used. However, the sample size for this response option labeling type was small, and the results must be interpreted with a grain of salt.

Discussion

This synthetic study set out to explore scale and study quality in L2 pronunciation research, focusing specifically on the measurement of three prominent constructs in the domain: accentedness, comprehensibility, and intelligibility. The study was guided by two RQs that examined the aspects of study quality followed by the aspects of instrument quality, including measurement instrument design, validity, and reliability. To address the RQs, a total of 380 articles containing 576 instruments were identified and coded for several variables of interest. Given that this is the first known principled synthesis of study and scale quality in L2 pronunciation research, we summarize and interpret the findings for each RQ contextualizing them within the larger field of applied linguistics.

Study quality

With regard to the first RQ, the findings are largely parallel to the state of study quality in applied linguistics overall and its subdomains of L2 pragmatics, language assessment, and individual differences (Kostromitina & Plonsky, 2022; Sudina, 2023; Taguchi, Kostromitina & Wheeler, 2022) as well as L2 pronunciation instruction itself (e.g., Lee et al., 2015). Specifically, a majority of primary studies in our sample recruited listeners and speakers among adult university students; however, there seems to be a move toward diversification of participants in recent years by sampling from Amazon Mechanical Turk workers and HR specialists as listeners and high school students, test

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

takers, and teachers as speakers. Another area of concern is that the location of participants was largely limited to English-centric contexts (e.g., universities in the US, the UK, and Canada). Moreover, the target language of investigation in over 70% of studies was English; unsurprisingly, it was also the predominant L1 of listeners. These potential shortcomings in participant sampling and recruitment as well as the focus on English to the detriment of other languages have been noted in previous synthetic studies in other areas of applied linguistics (Al-Hoorie, 2018; Sudina, 2021, 2023; Teimouri, Goetze & Plonsky, 2019) due to presenting the risk of making ungrounded generalizations about L2 learning and teaching. If certain groups are consistently overlooked in primary studies (e.g., older learners, children, and learners of languages other than English), it restricts our understanding of L2 pronunciation processes in these populations; researchers must recognize this limitation when interpreting their results (King & Mackey, 2016; Ortega, 2005). Contextualizing these findings in current methodological discussions surrounding L2 speech perception in particular, it is noteworthy that the majority of studies in our sample employed L1 as opposed to L2 listeners. It appears that, as a field, the focus on L1 listener judgments in evaluating L2 speakers' accentedness, comprehensibility, and intelligibility is still prevailing. In part, it may be the case that researchers tend to choose L1 listeners for their studies to avoid potential confounding effects of L2 accent familiarity and/or mutual intelligibility benefit (Bent & Bradlow, 2003) on the ratings. While we cannot make inferences about whether L1 or L2 listeners are preferred for perceptual judgments, we hope that more studies will examine the effect of listeners' L1 and the interaction between listener and speaker L1s in L2 speech ratings.

Some of the additional weaknesses in the quality of L2 pronunciation research include small sample sizes compared to L2 research overall. In our sample of primary studies, the median sample size was 25 for the speakers and 24 for the listeners. This is considerably smaller than the sample size in the larger domain of L2 research as reported by Plonsky (2014; median = 62 for studies between 2000 and 2009). While small samples are quite common in quantitative L2 research (Loewen & Hui, 2021), the size of a sample should be appropriate for the chosen statistical analyses; larger sample sizes tend to increase the statistical power of a study and are likely to present the population of interest more accurately. It is important to acknowledge, however, that in L2 pronunciation research, comprehensibility, accentedness, and intelligibility judgments are oftentimes paired with time-consuming and laborious acoustic speech analysis to examine, for example, which speech features predict such judgments (e.g., Kang, 2010; Kang et al., 2018). Therefore, smaller sample sizes, especially on the speaker side, might be warranted for research to be feasible.

Other aspects of study quality that are concerning include the lack of thorough reporting of participant demographic characteristics. That is, almost half of the studies in our sample did not report participants' gender, age, and L2 proficiency. It appears that failure to report basic information about the study sample found in synthetic research in the early 2010s (e.g., Plonsky, 2013; Plonsky & Gass, 2011) remains an issue to this day.

Instrument quality

Delving into the findings related to the second RQ, L2 pronunciation research showed drawbacks in latent construct operationalization, insufficient reporting of measurement instrument origin, development, and use as well as limited instrument validity

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

and reliability evidence. Our sample of primary studies included 30 investigations that measured intelligibility via a scale. This finding highlights the inconsistencies in operationalizing intelligibility and a tendency to conflate it with comprehensibility (as noted by Thomson, 2017). If anything, L2 pronunciation research demonstrates a lack of agreement on the operationalization of intelligibility, which persists years after Derwing and Munro (1997) published their seminal work attempting to unify the way accentedness, comprehensibility, and intelligibility were understood in the field. Importantly, in our sample of studies, we did not find instances of scale-based intelligibility measurement before Derwing and Munro's seminal publication. The field is yet to agree on a specific instrument used to measure intelligibility, perhaps due to the unique nature of this construct. Intelligibility goes beyond listeners' impression of ease of understanding and reflects the actual proportion or amount of speech being successfully understood. This definition of intelligibility may be the reason why this construct has been measured quite inconsistently across primary studies, which is reflected in the findings of this synthesis.

Nevertheless, it is encouraging to see that relative to the number of studies measuring accentedness and comprehensibility via a scale, the proportion of scale-based intelligibility-centered studies is rather small. This could indicate a move toward consistency in the operationalization of these constructs. Another muddling finding of the present synthesis is that all three constructs of interest were predominantly measured via a single-item instrument (i.e., a 1-item scale) except for 27 studies that used multiple items. While using a single scale may be an efficient way of collecting listener judgments, especially when it comes to naïve and untrained listeners, the complexity of L2 pronunciation constructs may require more than 1 item to tap into each of them more precisely. Several issues have been brought up in previous research on content validity related to using a single item to measure a latent construct. These issues are related to the difficulty of establishing reliability of a single-item scale as well as the potential underrepresentation of the construct of interest (DeVellis & Thorpe, 2022; Sudina, 2023). This approach is particularly problematic for measuring accentedness, comprehensibility, and intelligibility due to existing disagreements on their conceptualization and operationalization in the field. In addition to single-item instruments, item evaluation in the current sample of primary studies was performed for three instruments only, thereby suggesting additional gaps in content validity evidence.

To focus on methodological strengths and highlight meritorious reporting practices in the field, we will discuss effective item evaluation practices in more detail. For example, Aksakallı and Yağız (2020), in developing a rubric to measure comprehensibility, solicited feedback from five expert language instructors with "adequate knowledge of English pronunciation" (p. 15). Ibnian and Yaseen (2018) submitted a newly developed accentedness scale to a panel of university instructors in the fields of testing of English as a foreign language, phonology, and sociolinguistics; the changes proposed by the panel were accounted for when finalizing the scale. These studies are great examples of researchers taking care in developing the scales and collecting content validity evidence for them; unfortunately, these studies constitute less than 1% of the sample. It appears that other domains of applied linguistics demonstrate more promising trends in content validity analyses. For example, according to Sudina's (2021, 2023) methodological syntheses, anxiety, willingness to communicate, and motivation were rarely measured via a single-item scale (between 2% and 5% of all scales). Additionally, although item evaluation practices were still scant, there was a larger percentage of scales that were subject to an expert

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

review (from 10% for anxiety to 14% for willingness to communicate scales; Sudina, 2023) in comparison to the current findings on L2 pronunciation instruments. The evidence of construct validity of the instruments in this synthesis was also limited. We did not find studies where the instrument was tested for measurement invariance or convergent/discriminant validity. There were singular studies where authors provided an existing validity reference to a scale that had been previously validated (e.g., French, Gagné & Collins, 2020; Ruivivar & Collins, 2019; Saito & Saito, 2017; Sheredekina et al., 2022; Tsunemoto, Trofimovich & Kennedy, 2023). For example, Ruivivar and Collins (2019) stated that the 1,000-point scale used in their study was "adapted from Saito, Trofimovich and Isaacs (2017), and validated in other speech rating studies" (p. 277; specific citations were provided). Saito and Saito (2017) asked the readers to refer to a different article regarding scale validity. A few studies in the sample performed Rasch analysis on their instruments (e.g., Dalman & Kang, 2023; Shintani, Saito & Koizumi, 2019), which can be regarded as another type of construct validity evidence. Overall, limited construct validity evidence in scale-based L2 pronunciation research is congruent with underreported construct validity elsewhere in applied linguistics and psychological sciences (Flake & Fried, 2020; Sudina, 2023). Importantly, this gap in validity evidence does not only diminish the quality of reporting practices in L2 pronunciation research but also casts doubt on how accentedness, comprehensibility, and to some extent, intelligibility have been operationalized and whether the instruments measuring these constructs are sufficiently rigorous and reflective of the underlying latent constructs.

Regarding construct validity, the domain of L2 pronunciation could draw from validation frameworks that exist in language assessment (e.g., Knoch et al., 2021). However, an important distinction must be made between scales used for assessing proficiency and perceptual judgment scalar ratings used to measure accentedness, comprehensibility, and (sometimes) intelligibility. Because studies measuring the three constructs are typically interested in intuitive, impressionistic judgments, it might occasionally make sense to use single-item scales to simplify the rating process. Nevertheless, the latent nature of these constructs necessitates a clear operationalization, which involves identifying key components or dimensions established in previous research and accounting for the purpose of the measurement. For L2 comprehensibility, for example, one may consider specific segmental and suprasegmental features that have been found to contribute to the construct and develop scales based on these features. To illustrate, there were several studies in our sample (e.g., Dai & Roever, 2019; Kang, 2010) that operationalized L2 comprehensibility and accentedness via multi-item scales, providing more refined operationalizations of the constructs. For example, Kang (2010) used a 5-item semantic differential scale; each item measured the following aspects of comprehensibility: effort to understand, clarity, ease of understanding, ease of grasping the meaning, and overall comprehensibility. Using multi-item scales may also allow researchers to get a more fine-grained representation by measuring constructs related to L2 comprehensibility such as perceived fluency (Suzuki & Kormos, 2020); that is, one of the scale items could address specifically this construct. This does not mean that future research needs to avoid single-item scales. We are simply calling for more thought and consideration involved in operationalizing the constructs of interest.

Transparency was another quality feature that was largely missing from the studies in the current sample. We identified several gaps related to instrument availability (~72% of instruments were not available), pilot testing procedures (not reported for ~94% of instruments), authorship (not reported for 79% of instruments), and presence

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

of adaptations or lack thereof (not reported for ~87% of existing instruments). Descriptive statistics were not consistently reported either (mean scores were not reported for 26% of instruments; *SDs* were not reported for 45% of instruments). Similar problems with transparency were observed in L2 quantitative research by Plonsky (2013, 2014); however, the reporting of descriptive statistics in L2 pronunciation research appears to have improved relative to the trends found in L2 research in 2000–2010.

Regarding instrument design features, scales used in L2 pronunciation research appear to vary in terms of rating scale length, response format types, and response option labeling. More often than not, these scale design choices were not explained in the current sample of primary studies. That is, the studies in our sample often stated that a certain scale was used but never mentioned its author and origin (79% of instruments). In certain cases (e.g., Hansen, Zampini & Cunningham, 2019; Saito et al., 2020; Sučková, 2020; Tekin, 2019), scales were mislabeled as Likert(-style) scales even though the actual design was a semantic differential scale. As de Vaus (2016) explains, Likert scales usually include descriptors for each point of the scale (e.g., *strongly disagree*, *somewhat disagree*, *neutral*, *somewhat agree*, and *strongly agree*). Respondents then indicate their level of agreement, but the scale can also be used to indicate frequency, quality, likelihood, etc. Semantic differential scales, similar to the Likert ones, can still be numeric; however, they typically include opposing adjectives placed at the ends of the scale (p. 102). In cases of accentedness, comprehensibility, and intelligibility, these adjectives may include, for example, *very easy/hard to understand* (e.g., Kennedy & Trofimovich, 2008). The observed incorrect labeling of scales as Likert-(style) when they are, in fact, semantic differential scales is concerning as it can lead to methodological misunderstandings, which would in turn influence data interpretation and analysis. For one, Likert scales do not assume equal intervals between the scale points, and their analysis typically relies on the assumptions of ordinal data. In contrast, semantic differential scales are more flexible in capturing nuanced perceptions and can be analyzed assuming interval-level data (for a more technical discussion of semantic-differential scales, see Stoklasa, Talášek & Stoklasová, 2019). Therefore, mistakenly treating semantic differential scales as Likert ones may lead to inappropriate statistical analyses and invalid inferences, thereby affecting instrument and study quality. It also appears from the results of this synthesis that researchers may choose to use different response labeling options, most commonly partially verbal and numerical or partially verbal (endpoints only) in measuring the constructs of interest; however, these choices were rarely justified.

The final aspect of transparency and scale quality that we discuss in this section concerns instrument reliability; here, as with other aspects of study and scale quality, we found under-reporting. For multi-item scales, less than 10% (2 out of 27) reported item-total correlations. Moreover, almost half of the instruments did not have any interrater reliability reported. This is in line with the findings by Plonsky and Gass (2011) as well as Plonsky (2013, 2014) who observed 64%, 45%, and 50% of primary studies having reliability reported, respectively (although the present synthesis takes a more nuanced approach and examines reliability reporting practices at the scale level rather than at the study level). In fact, the rate of reliability reporting found in this study appears to be higher than that in other domains of applied linguistics, for example, as reported in a meta-analysis on the effects of instruction (Norris & Ortega, 2000) in SLA. For those instruments in our sample that had interrater reliability consistency reported, median agreement indices were relatively

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

high and independent of the type of scale used and the construct measured (see Table 7 in the Results section). The overall median interrater reliability of the instruments was .92 (IQR .14, k = 287[6]), which is comparable to estimates reported by Plonsky and Derrick (2016; median internal consistency = .82, median interrater reliability = .92), Sudina (2023; median internal consistency = .88), and Saito and Plonsky (2019; median interrater reliability ranging from .76 to .93).

When differentiated by moderator variables, RGM revealed that, in general, Cronbach's α and intraclass correlations in our study were relatively stable and independent of the rating scale length, type, and labeling preferences. The interrater reliability indices across these variables were consistent both for accentedness and comprehensibility with a few exceptions. Fully verbal scales appeared to show slightly lower interrater reliability compared to other types of response option labeling. We speculate that this observation may be due to the complexity of fully verbal scales for the untrained raters as well as the subjectivity related to the interpretation of responses (see also Menold & Bogner, 2016). At the rating scale length, all scales demonstrated high reliability estimates, with 9-point and 1,000-point scales having the most consistent estimates as measured by both Cronbach's α and intraclass correlation (see Table 7). This is possibly due to the higher number of these scales in the sample because a larger sample size may lead to higher reliability estimates due to the smaller margins of error (Kaye & Freedman, 2011). However, it is also not unlikely that these particular rating scale lengths contributed to listener raters' scoring consistency. This could be explained by increased variance of a scale with more points, leading to higher reliability coefficients, better discrimination, or reduced ceiling and floor effects (e.g., Lord & Novick, 2008). This finding supports other studies in the domain of L2 pronunciation that argued in favor of either 9-point scales (e.g., Munro, 2017) or 1,000-point scales for comprehensibility and accentedness measurement (e.g., Saito et al., 2017). At the same time, other studies in the field have argued for shorter scales (e.g., Isbell, 2017; Kermad & Bogorevich, 2022). Overall, it seems that L2 pronunciation research would benefit from greater consistency in the selection of scales that would facilitate the interpretation of results; critically, no matter the rating scale length, the choice of scale design needs to be motivated by research objectives and justified via reliability analyses.

Recommendations for researchers

Based on the findings in the current methodological synthesis as well as other synthetic literature on study quality, we would like to propose several recommendations for future studies in L2 pronunciation research that use perception-based judgments (noting, again, that intelligibility is a construct that is typically not measured with a scale).

Study quality and transparency

First, the field would benefit from an expansion of current research contexts and participant sampling outside of English-centric locations (see Plonsky, 2023). In general, investigations of accentedness, comprehensibility, and intelligibility need to

[6] The number of instruments used to calculate the overall median interrater reliability differs from the overall number of instruments with reported interrater reliability in Table 6. This is due to some studies reporting a range of reliability estimates rather than exact values (e.g., Behrman, 2014). Additionally, only scales that measured accentedness and comprehensibility were included in this calculation.

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

focus on languages other than English as the target. In terms of study transparency, the demographic characteristics of both listener and speaker participants need to be documented more thoroughly (e.g., gender, age, and target language proficiency for L2 speakers or listeners). Listening procedures (whether relistening to audio files is permitted, whether listening items are randomized, whether rating change is permitted, and whether and what kind of listener training is provided) need to be documented as part of the study design. Given the existing study design debates in L2 pronunciation (e.g., Kermad, 2021), it would be worthwhile to expand this line of research and provide clarity on the role of listeners' L1 in perception-based judgments. Researchers should also be clear in their definitions of listener training (i.e., considering one's background and experience or incorporating actual training procedures and calibration activities) to make more precise conclusions about its effects on the ratings.

Scale quality and transparency

Based on the findings of this synthesis, L2 pronunciation researchers collecting perceptual judgments about constructs like accentedness and comprehensibility might consider the following recommendations:

1. Constructs measured by the scales need to be rooted in and defined by previous research. In other words, consider how theoretical, framework-setting studies operationalized a construct.
2. The choice of an instrument to measure a construct needs to be motivated (e.g., the use of transcription versus a scale for intelligibility).
3. The steps in scale development or adaptation need to be documented and reported systematically. The following questions may be considered for documentation purposes: (a) Is this a newly developed scale or an adaptation? (b) If it is an adaptation, who are the authors, and what changes (if any) have been made to the original version? and (c) Was there any piloting done during scale development? If not, why?
4. The scale(s) should be made available in the study report itself, in an appendix, in supplementary materials online, or in a research database (e.g., IRIS; https://www.iris-database.org/) to help readers independently evaluate instrument and study quality, adhere to principles of open science, and enable future replications (McKay & Plonsky, 2021).
5. Scale design choices should be reported thoroughly and systematically, explaining and justifying each of the following: rating scale length, presence or absence of a neutral midpoint, scale type, and response option labeling.
6. Descriptive statistics should be reported for each scale employed in the study along with (a) content and construct validity evidence (e.g., by referring to a previous scale validation study if the instrument was used without major adaptations and with the same population) and (b) interrater reliability and/or internal-consistency reliability.
7. In terms of scale length, given the results of the reliability generation meta-analysis conducted as part of this study, the following scale features seem to lead to higher interrater reliability when measuring accentedness and comprehensibility: a 9-point or 1,000-point scale and anything but fully verbal labeling (although these other scale design features need to be further investigated in future studies on study quality). Although researchers may consider using these scale lengths over others to measure these two constructs, the ultimate length of the scale should be motivated

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

by additional aspects of study design like listener factors (e.g., age, availability of training) as well as the study goals.

Conclusion and future directions

The goal of this synthetic study was to examine study and scale quality in L2 pronunciation research, while focusing on the measurement of accentedness, comprehensibility, and intelligibility. To our knowledge, this was the first inquiry into perception-based quantitative L2 pronunciation research with a focus on study and instrument quality. The findings in this synthesis are important as they lay a foundational groundwork for future research endeavors in L2 pronunciation assessment. By highlighting the gaps in study and scale quality related to reporting practices, validity, and reliability evidence, this study underscores the need for more consistency, transparency, and overall methodological rigor in the domain.

The study presents several future research opportunities related to its limitations. For one, future research could expand the current findings by including non-articles and unpublished research (especially in languages other than English) as well as examining constructs and instrument types other than the ones chosen in this synthesis (e.g., L2 fluency). Given that some of the moderator analyses had to be conducted with relatively small samples, expanding the sample of studies included in the synthesis, especially its meta-analytic part, would provide ground for more generalizable findings. Specifically, we encourage synthetic researchers to expand on study design aspects that were outside of the scope of the current systematic review and RGM. It would be worthwhile to examine listener factors more closely, including the moderating effect of listener training along with its type and listener L1 background. Future studies could also examine features related to the sound files used for judgments (e.g., length of files, their quality, and word versus sentence versus text-level judgments) and how they may affect reliability of listener ratings. Finally, because the study and instrument quality have not been examined extensively in L2 pronunciation, future replications of this study could shed more light on this issue leading, in turn, to more robust research in the domain.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S027226312500018X.

Data availability statement. The experiment in this article earned Open Materials badges for transparent practices. The materials are available at https://doi.org/10.48316/10C5S-64Ug7

Acknowledgements. We would like to thank our handling editor Dr. Kazuya Saito, as well as the four anonymous reviewers for their helpful comments and suggestions. We also wish to thank Dr. Luke Plonsky and everyone who attended our talk and provided feedback on our study at AAAL-2024 in Houston, TX. Finally, we are grateful to Shelby Hans, a graduate research assistant, for her help with article searches and screening as well as Hande Özdemir for helping proofread and finalize the manuscript.

References

Aksakallı, C., & Yağız, O. (2020). The pre-service EFL teachers' development of phonological processing and evaluation of their attitudes toward pronunciation. *GIST - Education and Learning Research Journal*, *20*, 7–31. https://doi.org/10.26817/16925777.712

Al-Hoorie, A. H. (2018). The L2 motivational self system: A meta-analysis. *Studies in Second Language Learning and Teaching*, *8*, 721–754. https://doi.org/10.14746/ssllt.2018.8.4.2

Anthony, L. (2020). *AntConc* [Computer software]. Version 3.5.9. Waseda University. https://www.laurenceanthony.net/software

Baills, F., Zhang, Y., Cheng, Y., Bu, Y., & Prieto, P. (2021). Listening to songs and singing benefitted initial stages of second language pronunciation but not recall of word meaning. *Language Learning*, *71*, 369–413. https://doi.org/10.1111/lang.12442

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 54–74. https://doi.org/10.1080/15434300903464418

Behrman, A. (2014). Segmental and prosodic approaches to accent management. *American Journal of Speech-Language Pathology*, *23*(4), 546–561. https://doi.org/10.1044/2014_AJSLP-13-0074

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*, 1600–1610. https://doi.org/10.1121/1.1603234

Burda, A. N., & Hageman, C. F. (2005). Perception of accented speech by residents in assisted-living facilities. *Journal of Medical Speech-Language Pathology*, *13*(1), 7–15. https://link.gale.com/apps/doc/A131004085/HRCA?u=anon~8c3e5c44&sid=googleScholar&xid=832636b8

Byrnes, H. (2013). Positioning writing as meaning-making in writing research: An introduction. *Journal of Second Language Writing*, *22*, 95–106. https://doi.org/10.1016/j.jslw.2013.03.004

Chapelle, C.A. (2013). Reliability in language assessment. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 4918–4923). Blackwell/Wiley. http://works.bepress.com/carol_chapelle/29/

Chau, T., Huensch, A. (2025). The relationships among L2 fluency, intelligibility, comprehensibility, and accentedness: A meta-analysis. *Studies in Second Language Acquisition*, 1–26. Advance online publication. https://doi.org/10.1017/S0272263125000014

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. *TESL-EJ*, *25*, 1–29. https://doi.org/10.55593/ej.25100a7

Christiner, M., Bernhofs, V., Sommer-Lolei, S., & Groß, C. (2023). What makes a foreign language intelligible? An examination of the impact of musical ability and individual differences on language perception and how intelligible foreign languages appear. *Journal of Intelligence*, *11*, 43. https://doi.org/10.3390/jintelligence11030043

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

Dai, D. W., & Roever, C. (2019). Including L2-English varieties in listening tests for adolescent ESL learners: L1 effects and learner perceptions. *Language Assessment Quarterly*, *16*, 64–86. https://doi.org/10.1080/15434303.2019.1601198

Dalman, M., & Kang, O. (2023). Validity evidence: Undergraduate students' perceptions of TOEFL iBT high score spoken responses. *International Journal of Listening*, *37*, 113–126. https://doi.org/10.1080/10904018.2021.1929993

de Vaus, D. (2016). *Surveys in social research* (6th ed.). Routledge.

Derwing, T. M. (2008). Curriculum issues in teaching pronunciation to second language learners. *Phonology and Second Language Acquisition*, *36*, 347–369. https://doi.org/10.1075/sibil.36.17der

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. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*, 379–397. https://doi.org/10.2307/3588486

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, *29*, 359–380. https://doi.org/10.1093/applin/amm041

DeVellis, R. F. & Thorpe, C. T. (2022). *Scale development: Theory and applications* (5th ed.). Sage.

Ellis, J. L. (2021). A test can have multiple reliabilities. *Psychometrika*, *86*(4), 869–876. https://doi.org/10.1007/s11336-021-09800-2

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*, 456–465. https://doi.org/10.1177/2515245920952393

Foote, J. A., & Trofimovich, P. (2018). Is it because of my language background? A study of language background influence on comprehensibility judgments. *Canadian Modern Language Review*, *74*, 253–278. https://doi.org/10.3138/cmlr.2017-0011

French, L. M., Gagné, N., & Collins, L. (2020). Long-term effects of intensive instruction on fluency, comprehensibility and accentedness. *Journal of Second Language Pronunciation*, *6*, 380–401. https://doi.org/10.1075/jslp.20026.fre

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, *54*, 245–258. https://doi.org/10.1017/S0261444819000430

Ghanem, R., Kang, O., & Kostromitina, M. (2024). *L2 spoken discourse: Linguistic features and analyses*. Taylor & Francis. https://doi.org/10.4324/9780429030024

Gooch, R., Saito, K., & Lyster, R. (2016). Effects of recasts and prompts on L2 pronunciation development: Teaching English/ɹ/to Korean adult EFL learners. *System*, *60*, 117–127. https://doi.org/10.1016/j.system.2016.06.007

Hansen Edwards, J. G., Zampini, M. L., & Cunningham, C. (2019). Listener proficiency and shared background effects on the accentedness, comprehensibility and intelligibility of four varieties of English. *Journal of Monolingual Bilingual Speech*, *1*, 333–356. https://doi.org/10.1558/jmbs.v1i2.11867

Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency*. John Benjamins. https://doi.org/10.1075/lllt.32

Ibnian, S. S., & Yaseen, M. S. (2018). The level of fluency in articulating American English among Jordanian Arabic speakers. *European Journal of Scientific Research*, *150*, 201–212.

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

Isaacs, T., & Trofimovich, P. (Eds.). (2016). *Second language pronunciation assessment: Interdisciplinary perspectives* (Vol. 107). Multilingual Matters. https://doi.org/10.21832/ISAACS6848

Isaacs, T., Trofimovich, P., & Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, *35*, 193–216. https://doi.org/10.1177/0265532217703433

Isaacs, T., Trofimovich, P., Yu, G., & Chereau, B. M. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS Research Reports Online Series*. https://ielts.org/researchers/our-research/research-reports/examining-the-linguistic-aspects-of-speech-that-most-efficiently-discriminate-between-upper-levels-of-the-revised-ielts-pronunciation-scale

Isbell, D. R. (2017). Assessing pronunciation for research purposes with listener-based numerical scales. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 89–112). Routledge. https://doi.org/10.4324/9781315170756-6

JASP Team (2024). *JASP* [Statistical software]. Version 0.18.3. https://jasp-stats.org

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, *38*, 301–315. https://doi.org/10.1016/j.system.2010.01.005

Kang, O., & Rubin, D. (2012). Intra-rater reliability of oral proficiency ratings. *The International Journal of Educational and Psychological Assessment*, *12*, 43–61.

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

Kaye, D. H., & Freedman, D. (2011). *Reference guide on statistics*. National Academy Press. https://ssrn.com/abstract=2705655

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

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

Kermad, A. (2021). Training the "everyday" listener how to rate accented speech. *International Journal of Listening*, *38*, 58–78. https://doi.org/10.1080/10904018.2021.1987910

Kermad, A., & Bogorevich, V. (2022). Using statistical transformation methods to explore speech perception scale lengths. *Language Teaching Research Quarterly*, *29*, 65–91. https://doi.org/10.32038/ltrq.2022.29.05

King, K. A., & Mackey, A. (2016). Research methodology in second language studies: Trends, concerns, and new directions. *Modern Language Journal*, *100*, 209–227. https://doi.org/10.1111/modl.12309

Kissling, E. M. (2014). What predicts the effectiveness of foreign-language pronunciation instruction? Investigating the role of perception and other individual differences. *Canadian Modern Language Review*, *70*, 532–558. https://doi.org/10.3138/cmlr.2161

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing*, *38*(4), 602–626. https://doi.org/10.1177/0265532221994052

Kostromitina, M., & Plonsky, L. (2022). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, *44*, 886–911. https://doi.org/10.1017/S0272263121000395

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

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*, 369–377. https://doi.org/10.2307/3588485

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

Levis, J. (2020a). Changes in L2 pronunciation: 25 years of intelligibility, comprehensibility, and accentedness. *Journal of Second Language Pronunciation*, *6*, 277–282. https://doi.org/10.1075/jslp.20054.lev

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

Loewen, S., & Hui, B. (2021). Small samples in instructed second language acquisition research. *Modern Language Journal*, *105*, 187–193. https://doi.org/10.1111/modl.12700

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. Information Age Publishing Inc.

Matsuura, H., Chiba, R., Mahoney, S., & Rilling, S. (2014). Accent and speech rate effects in English as a lingua franca. *System*, *46*, 143–150. https://doi.org/10.1016/j.system.2014.07.015

McKay, T., & Plonsky, L. (2021). Reliability analyses: Estimating error in L2 research. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 468–482). Routledge.

McKay, T. H., Teimouri, Y., Sağdıç, A., Salen, B., Reagan, D., & Malone, M. E. (2021). The cagey C-test construct: Some evidence from a meta-analysis of correlation coefficients. *System*, *99*, 102526. https://doi.org/10.1016/j.system.2021.102526

Menold, N., & Bogner, K. (2016). *Design of rating scales in questionnaires* (Version 2.0). (GESIS Survey Guidelines). GESIS - Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.15465/gesis-sg_en_015

Meritan, C. (2022). Integrating online pronunciation instruction: The case of learners of French. *Foreign Language Annals*, *55*, 877–893. https://doi.org/10.1111/flan.12646

Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and Second Language Acquisition* (pp. 193–218). John Benjamins. https://doi.org/10.1075/sibil.36.10mun

Munro, M. J. (2017). Dimensions of pronunciation. In O. Kang, R.I. Thomson, & J.M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 413–431). Routledge. https://www.routledgehandbooks.com/doi/10.4324/9781315145006-4

Munro, M. J., & Derwing, T. M. (1995). 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. (2015). Intelligibility in research and practice: Teaching priorities. 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., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in second language acquisition*, *28*(1), 111–131. https://doi.org/10.1017/S0272263106060049

Nadler, J. T., Weston, R., & Voyles, E. C. (2015). Stuck in the middle: The use and interpretation of midpoints in items on questionnaires. *The Journal of General Psychology*, *142*, 71–89. https://doi.org/10.1080/00221309.2014.994590

Nagle, C. L., & Baese-Berk, M. M. (2022). Advancing the state of the art in L2 speech perception-production research: Revisiting theoretical assumptions and methodological practices. *Studies in Second Language Acquisition*, *44*, 580–605. https://doi.org/10.1017/S0272263121000371

Norouzian, R. (2021). Interrater reliability in second language meta-analyses: The case of categorical moderators. *Studies in Second Language Acquisition*, *43*, 896–915. https://doi.org/10.1017/S0272263121000061

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417–528. https://doi.org/10.1111/0023-8333.00136

Norris, J. M., & Ortega, L. (2012). Assessing learner knowledge. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 573–589). Routledge.

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

Ortega, L. (2005). For what and for whom is our research? The ethical as transformative lens in instructed SLA. *Modern Language Journal*, *89*, 427–443. https://doi.org/10.1111/j.1540-4781.2005.00315.x

Plonsky, L. (n.d.). *Second-language Research Corpus*. Unpublished database.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687. https://doi.org/10.1017/S0272263113000399

Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal*, *98*, 450–470. https://doi.org/10.1111/j.1540-4781.2014.12058.x

Plonsky, L. (2023). Sampling and generalizability in Lx research: A second-order synthesis. *Languages*, *8*, 75–88. https://doi.org/10.3390/languages8010075

Plonsky, L. (2024). Study quality as an intellectual and ethical imperative: A proposed framework. *Annual Review of Applied Linguistics*, 1–15. https://doi.org/10.1017/S0267190524000059

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

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366. https://doi.org/10.1111/j.1467-9922.2011.00640.x

Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research*, *36*, 583–621. https://doi.org/10.1177/0267658319828413

Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). Routledge. https://doi.org/10.4324/9781315870908-6

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … & Moher, D. (2021). Updating guidance for reporting systematic reviews: Development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology*, *134*, 103–112. https://doi.org/10.1016/j.jclinepi.2021.02.003

Raaijmakers, Q. A. W., van Hoof, J. T. C., 't Hart, H., Verbogt, T. F. M. A., & Vollebergh, W. A. M. (2000). Adolescents' midpoint responses on Likert-type scale items: Neutral or missing values. *International Journal of Public Opinion Research*, *12*, 208–216. https://doi.org/10.1093/ijpor/12.2.209

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review*, *65*(3), 395–412. https://doi.org/10.3138/cmlr.65.3.395

Ruivivar, J., & Collins, L. (2019). Nonnative accent and the perceived grammaticality of spoken grammar forms. *Journal of Second Language Pronunciation*, *5*, 269–293. https://doi.org/10.1075/jslp.17039.rui

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., Ilkan, M., Magne, V., Tran, M. N., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid-and high-level second language fluency. *Applied Psycholinguistics*, *39*, 593–617. https://doi.org/10.1017/S0142716417000571

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., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech? Roles of first language profiles, second language proficiency, age, experience, familiarity, and metacognition. *Studies in Second Language Acquisition*, *41*, 1133–1149. https://doi.org/10.1017/S0272263119000226

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

Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2016). Re-examining phonological and lexical correlates of second language comprehensibility: The role of rater experience. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 141–156). Multilingual Matters. https://doi.org/10.21832/ISAACS6848

Saito, Y., & Saito, K. (2017). Differential effects of instruction on the development of second language comprehensibility, word stress, rhythm, and intonation: The case of inexperienced Japanese EFL learners. *Language Teaching Research*, *21*, 589–608. https://doi.org/10.1177/1362168816643111

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press

Saito, K., Trofimovich, P., Abe, M., & In'nami, Y. (2020). Dunning-Kruger effect in second language speech learning: How does self perception align with other perception over time? *Learning and Individual Differences*, *79*, 101849. https://doi.org/10.1016/j.lindif.2020.101849

Sheredekina, O., Karpovich, I., Voronova, L., & Krepkaia, T. (2022). Case technology in teaching professional foreign communication to law students: Comparative analysis of distance and face-to-face learning. *Education Sciences*, *12*, 645. https://doi.org/10.3390/educsci12100645

Shintani, N., Saito, K., & Koizumi, R. (2019). The relationship between multilingual raters' language background and their perceptions of accentedness and comprehensibility of second language speech. *International Journal of Bilingual Education and Bilingualism*, *22*, 849–869. https://doi.org/10.1080/13670050.2017.1320967

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Stoklasa, J., Talášek, T., & Stoklasová, J. (2019). Semantic differential for the twenty-first century: Scale relevance and uncertainty entering the semantic space. *Quality & Quantity*, *53*, 435–448. https://doi.org/10.1007/s11135-018-0762-1

Sučková, M. (2020). Acquisition of a foreign accent by native speakers of English living in the Czech Republic. *ELOPE: English Language Overseas Perspectives and Enquiries*, *17*, 83–100. https://doi.org/10.4312/elope.17.2.83-100

Sudina, E. (2021). Study and scale quality in second language survey research, 2009–2019: The case of anxiety and motivation. *Language Learning*, *71*, 1149–1193. https://doi.org/10.1111/lang.12468

Sudina, E. (2023). Scale quality in second-language anxiety and WTC: A methodological synthesis. *Studies in Second Language Acquisition*, *45*, 1427–1455. https://doi.org/10.1017/S0272263122000560

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

Taguchi, N., Kostromitina, M. M., & Wheeler, H. (2022). Individual difference factors for second language pragmatics. In S. Li, P. Hiver, & M. Papi (Eds.), *The Routledge handbook of second language acquisition and individual differences* (pp. 310–330). Routledge. https://doi.org/10.4324/9781003270546-26

Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, *41*, 363–387. https://doi.org/10.1017/S0272263118000311

Tekin, O. (2019). The association between ethnic group affiliation and the ratings of comprehensibility, intelligibility, accentedness, and acceptability. *The Electronic Journal for English as a Second Language*, *24*, 1–29.

Thomson, R. (2017). Measurement of accentedness, intelligibility, and comprehensibility. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 11–29). Routledge. https://doi.org/10.4324/9781315170756-2

Tsunemoto, A., Lindberg, R., Trofimovich, P., & McDonough, K. (2022). Visual cues and rater perceptions of second language comprehensibility, accentedness, and fluency. *Studies in Second Language Acquisition*, *44*, 659–684. https://doi.org/10.1017/S0272263121000425

Tsunemoto, A., Trofimovich, P., & Kennedy, S. (2023). Pre-service teachers' beliefs about second language pronunciation teaching, their experience, and speech assessments. *Language Teaching Research*, *27*, 115–136. https://doi.org/10.1177/1362168820937273

Yan, X., & Ginther, A. (2017). Listeners and raters: Similarities and differences in evaluation of accented speech. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 67–88). Routledge. https://doi.org/10.4324/9781315170756-5

Cite this article: Kostromitina, M., Sudina, E., & Baghlaf, E. (2025). Study and instrument quality in perception-based L2 pronunciation research: A methodological synthesis. *Studies in Second Language Acquisition*, 1–34. https://doi.org/10.1017/S027226312500018X

https://doi.org/10.1017/S027226312500018X Published online by Cambridge University Press