How many raters should be enough: G Theory Applied to Assessment and Measurement of L2 Speech Perception

Kevin Hirschi, Okim Kang

Department of English, Northern Arizona University

Author Note

Correspondence concerning this article should be addressed to Kevin Hirschi, Department of English, Northern Arizona University, P.O. Box 6032, Flagstaff, AZ 86011, United States. Email: KevinHirschi@nau.edu

Abstract

This paper extends the use of Generalizability Theory to the measurement of extemporaneous L2 speech through the lens of speech perception. Using six datasets of previous studies, it reports on *G studies*–a method of breaking down measurement variance–and *D studies*–a predictive study of the impact on reliability when modifying the number of raters, items, or other facets, that assists the field in adopting measurement designs that include comprehensibility, accentedness, and intelligibility. When data of a single audio sample from a learner were subjected to D-studies, we find that both semantic differential and rubric scales for comprehensibility were reliable at the .90 level with about 15 trained raters or 50 untrained crowdsourced raters. In order to offer generalizable and dependable evaluations, empirically informed recommendations are

given, including considerations for number of speech samples rated, or the granularity of the scales for various assessment and research purposes.

Keywords: generalizability, L2 speech, comprehensibility, accentedness, intelligibility, measurement

In recent years, pedagogical targets for L2 learners have pivoted away from native-like speech and move towards intelligibility and comprehensibility in pronunciation (Levis, 2020). Concomitantly, teachers and researchers alike have sought accurate and meaningful approaches to measure L2 speech through more holistic measures based on perception by experts or lay listeners. Indeed, the impacts of Munro and Derwing's (1995) description of comprehensibility, intelligibility, and accentedness have influenced testing, classroom, intervention, and correlational research (For a review, see Levis, 2020; Saito, 2021; Saito & Plonsky, 2019). These global constructs have been operationalized with a variety of measurement approaches largely centered around several listeners listening to different speakers and rating them on a scale.

In efforts to refine measurement, researchers have investigated the use of different scale lengths (Isaacs & Thomson, 2013; Kermad & Bogorevich, 2022), rater backgrounds (Saito, 2021; Saito et al., 2016; Saito & Shintani, 2016), and rater training (Authors and Colleagues., 2019; Colleagues in Authors, 2017), finding evidence for the validity of some scale lengths with different combinations of raters and types of training. However, none has systematically investigated the impact of the number of raters on speech evaluation and perceptual outcomes in L2 speech research.

This study explores the number of raters required to attain generalizable and dependable measurement for mostly widely researched L2 speech constructs, i.e., comprehensibility, accentedness, and intelligibility of extemporaneous speech using the Generalizability (G) Theory framework (Brown, 2013; Cronbach et al., 1963, 1972; Shavelson & Webb, 1991). Six datasets of L2 speech ratings were analyzed by using G Theory to shed light on the number of raters required for generalizable and dependable measurement and analyses of sources of variance in studies with different scale lengths and types of raters (i.e., trained linguists vs. naive listeners). Through additional steps within the G Theory framework, predictions of the modification of the number of raters were computed, resulting in empirically informed recommendations for L2 speech rating designs to enhance measurement practices.

The Review of Literature

L2 Speech Perception Rating

Throughout the past three decades, L2 speech research has experienced a paradigm shift away from purely acoustic measurements of speech to adoption of Munro and Derwing's (1995) listener-focused constructs of comprehensibility, intelligibility, and accentedness. Comprehensibility, according to Derwing and Munro's (1997) seminal work, consists of "judgments on a rating scale of how difficult or easy an utterance is to understand" (p. 2). Inherent in the conceptualization of comprehensibility is the use of a scale and multiple listeners evaluating one speech sample. In efforts to distinguish comprehensibility from foreign accented speech, listeners have also been asked to evaluate the magnitude of a foreign accent as a related but separate phenomenon, also using a rating scale. Intelligibility, on the other hand, refers as how much "a speaker's message is actually understood" (Munro & Derwing, 1995, p. 76), and typically involves computing a percentage of correctly transcribed words by several listeners.

Rating designs have ranged from the use of 9-point scales for accentedness and comprehensibility amongst listener samples as small as three applied linguists in Trofimovich et al. (2016), to 82 listeners with a range of experience with linguistics but limited L2 Korean exposure and "varying levels of experience with L2 speakers" (Isbell & Lee, 2022, p. 815). More recently, Miao et al., (2023) implemented a listener panel design with 687 listeners, in mixed groups of about 120 of both L1 and L2, as well as expert and naïve listeners. Generally, researchers have used fewer listeners if the listeners were trained or had a background in linguistics. For example, Crowther et al. (2015) asked 10 current or recent graduate students of Applied Linguistics to rate comprehensibility using a free-moving 1000-point slider scale is used. In designs with more listeners, Huensch and Nagle (2021) crowdsourced ratings using 80 L1 Spanish raters on Mechanical Turk to evaluate accentedness and comprehensibility using 100-point sliding scales, accompanied with a transcription task. Studies that included students or learners as listeners included between 25 and 60 listeners, including Munro & Derwing (2001)'s Study 1 with 48 university monolingual English students or Study 2 with 27 educational psychology students. These two listener groups used 9-point scales for accentedness and comprehensibility. Author (2010) included 58 undergraduate students with varied L2 experience. The scales in this study were 7-point scales and included multiple items for each construct. To this end, Saito et al. (2017) compared Cronbach's alpha coefficients between 10 inexperienced and 10 experienced raters using also using an unlabeled 1000-point slider scale, finding high reliability for comprehensibility and ($.94 < \alpha < .95$) even higher ($.95 < \alpha < .97$) for accentedness. However, it should be noted that the speech samples represented a wide range of speaking abilities, and all came from the same L1 English background with high familiarity of the L2, French.

Intelligibility of extemporaneous speech, on the other hand, is commonly measured by asking listeners to transcribe the L2 speech they hear, often after a single listen. These transcriptions are then compared to a more carefully constructed transcript when using extemporaneous speech or the original text in the case of scripted speech. Foundational to the transcription comparison approach is the extensive analysis conducted by Munro and Derwing (1995) in which function, content, word addition or deletion, or regularizations were tabulated separately from 18 introductory linguistics university students of 36 speech samples. Overall, the number of accurately transcribed words served as an informative variable in comparison to the other speech constructs they measured. Recently, Huensch and Nagle (2021) engaged a larger number of listener transcribers (80 MTurk) following Munro and Derwing's (1995) coding scheme on 42 L2 speech samples. While the results of these two studies were informative in terms of the relationships amongst the constructs, reliability or coder agreement is not available from these designs. Kennedy and Trofimovich (2008) describe a process of scoring correctly transcribe content words only amongst 24 listeners, half of which were naïve listeners with minimal exposure to L2 accents and the other half were teachers. Their resulting reliability coefficients ranged from .72 to .84, depending on the group of listeners. In sum, the field has yet to agree on use of reliability analyses of intelligibility transcription tasks.

Inherent in these designs is the assumption of homogeneity of listener abilities to decode L2 speech. Whether ratings are modeled as a single mean, or a random effect in a mixed effect model (Nagle, 2018), research questions and statistical analyses reinforce the expectation that listeners may react consistently across different examples of L2 speech. Isbell's (2017) response pattern and Rasch analysis of ten listeners evaluating comprehensibility on a 9-point scale indicated that raters are more consistent when evaluating comprehensibility than accentedness,

but demonstrate inconsistencies amongst groups of raters and question whether the 9-point scale results in points that are truly equidistant. Most pertinent to the present study is the meta-analysis of rater reliability conducted by Saito (2021) in which reliability coefficients were compared between novice, expert, and L2 listeners of accentedness and comprehensibility from 57 listener panels in previous published research. Coefficients were generally high (> .87) across all synthesized reliability coefficients with minimal variance between types of raters or target construct. Given the wide range of designs in terms of the number of listeners, speech samples, and scale lengths, researchers are faced with questions of logistics and informed guesses when planning listener panel studies as they seek to consider the generalizability of the scores in the real-world contexts. Furthermore, researchers must balance the monetary and time cost of large-scale listener designs while simultaneously seeking numerical accuracy. In order to investigate these questions, we turn to *Generalizability Theory*.

Generalizability Theory

Generalizability (G) Theory was codified by Cronbach et al. (1972) as a supplemental approach to classical test theory that allows for investigation of the error associated with psychometric measurement across *facets* (i.e., sources of error) such as rater, task, scale length, and subject. It is differentiated from classical theory in terminology in that it does not refer to measurement reliability, rather it examines (a) *generalizability* for norm-referenced tests, (b) *decision dependability* for test results that are used as cut points for grouping participants, and (c) the *dependability* for criterion-referenced tests (Brown, 2013). It also can provide estimations of error of increasing or decreasing the number of raters, items, categories, or other facets of the rating design.

Statistically, G Theory parallels the factorial ANOVA approach to explanation of variance in that it seeks to determine the extent to which a set of ratings can be generalized to a universe of ratings. In fact, a commonly used software to compute GT data is dubbed GENOVA (Crick & Brennan, 1982). G Theory proposes two possible designs to assist the understanding of error derived from psychometric designs: a *Generalizability* (G) study and a *Decision* (D) study.

A Generalizability (G) study determines the variance of each facet within the rating scheme (Shavelson & Webb, 1991). For example, a rating process might have a number of raters, persons (i.e., subjects), and multiple tasks per person. The G study computes variance for each of these facets and ratios of variance per facet in percentages. In ideal settings, the *persons* facet (i.e., individual's ability as approximated by a score) explains the most variance if the sample is heterogenous. However, it may be the case that raters were poorly trained or that instruments can be interpreted in unexpected ways, both contributing to the variance more than the inter-person variability. A Decision (D) Study, on the other hand, generates a generalizability coefficient and a dependability coefficient, both of which are on a scale of 0 to 1 (see Brown, 2013). In a norm-referenced test, the generalizability coefficient represents the overall generalizability to a universe of scores of the entire scoring procedure and is analogous to reliability in classical testing. In criterion-referenced testing contexts, the dependability coefficient is used, which represents the replicability of the scoring procedure that has been conducted. The difference in the dependability coefficient is that it statistically accounts for the potential skewness before or after an intervention found in criterion-referenced testing. Classical test procedures (e.g., Cronbach's alpha, Intra Class Correlation) do not (Shavelson & Webb, 1991).

Additionally, the D study can estimate changes in the rating design using the known generalizability and dependability coefficients. Relatively simple math that includes the error variance allows a researcher to determine the estimated impacts on generalizability if levels within the facets are reduced or expanded. Again, taking the example of raters, persons, and tasks, the D study can create a table or scatterplot that answers the question *what happens to my generalizability / dependability if I had more / fewer raters or tasks*? Given these estimated coefficients, administrators can design their rating for a fully crossed or partially crossed rating design to reduce the resources necessary to carry out the rating.

G Theory in L2 Speech Research

Several notable studies have contributed to the field of language teaching and testing using G Theory approaches to generalizability and dependability. In L2 pragmatics research, Brown and Ahn (2011) compared the generalizability of different types of Discourse Completion Tasks (DCTs). In this study, oral, written, role-play, and self-assessed DCTs were compared, finding minimal differences in generalizability and dependability between the DCT types, but that role plays were slightly less influenced by interactions of rating facets. Focusing on variance introduced by raters, Brown and Ahn (2011) found that the rater alone accounted for 1.19% to 3.29% of the rating variance. One other notable use of G Theory comes from idiomatic expression research. Hubers et al. (2019) employed subjective ratings of idiom frequency, usage, familiarity, and comprehension using a 5-point scale, open response, or multiple choice response. The D study results provided insights into the number of raters needed for idiomatic expression norming that cannot be calculated with classical reliability measures.

Relatively few studies within L2 speech appear to have employed G Theory to assess their rating procedures, perhaps because of its traditional use within assessment literature. Kim

(2009) conducted a G study on L1 and L2 raters of spoken proficiency. The results were similar between the two types of raters (0.0% to 0.2% of the variance was explained by the raters alone) and provided insights into potential efficiencies that could be made in the rating design. In a pedagogically-focused review of the impact of suprasegmental pronunciation feature recognition and production, Gorsuch (2001) used G Theory to determine the scale dependability of a post-test in a pre- / post- design. She found that the scale was not dependable with a generalizability coefficient of .45 (G study) and coefficient of .61 in the D study would be reached if there were 12 tasks and 2 raters. Turning to high-stakes testing context of International Teaching Assistants, Shin (2022) employed G Theory to investigate potential L1 bias. The results, however, indicate that the ratings did not reflect linguistic bias, and that fewer items could result in similarly high dependability with additional modifications. Similar empirically informed rating designs have yet to be employed for more global constructs such as comprehensibility.

However, research within speech pathology, which often uses a similar transcription task for intelligibility or visual-analogue scales have employed approaches grounded in G Theory to inform the effects of rater design on reliability. Xue et al. (2023) found that five expert raters were sufficient to provide reliable rating-based and transcription-based measures for scripted dysarthric speech (e.g., speech impacted by Parkinsons, Multiple Sclerosis, Congenital) and that reliability increased as the number of raters or utterances increased. Their study further shed light on the advantages brought about by adding an additional utterance (i.e., two speech samples from the same speaker at the same time) when studying generalizability in a multivariate D study design. Still, very little research has applied G Theory to L2 perception research in order to validate the number of raters needed to establish sufficient reliability of the study.

The Present Study

Rating designs for L2 comprehensibility measurement have been investigated in comparative forms based on rater training (Saito, 2021; Xi & Mollaun, 2009; Colleagues, 2017), suggesting the need for a differentiated approach to selecting the number of raters based on their experience and training rigor. However, none of these studies have directly investigated the number of raters needed to achieve accurate scoring. Furthermore, little is known about the potential impact on rater variance given different scale lengths (i.e., 5-point, 9-point, 100-point, 1000-point). To this end, this study investigates the generalizability and dependability of L2 speech, perception variables with L2 speech constructs (i.e., comprehensibility, accentedness, and intelligibility), considering rater training and scale length. Note the measurement of intelligibility is especially controversial in the field due to entrenched measurement approaches (Authors and Colleagues, 2018). The current intelligibility measurement approach is very exploratory, and the interpretation of the findings should be made in a limited context. The study is guided by the following three research questions:

RQ1. To what extent are rating designs for L2 comprehensibility generalizable and dependable at different scale lengths and rater backgrounds?

RQ2. To what extent are rating designs for L2 accentedness generalizable and dependable at different scale lengths and rater backgrounds?

RQ3. To what extent are transcription task scores for L2 intelligibility generalizable and dependable?

Method

Datasets

Six datasets from three existing studies were chosen to investigate generalizability and dependability of listener ratings of comprehensibility, accentedness, and intelligibility sourced from Authors and Colleagues (2023a), Authors and Colleagues (2020), and Authors and Colleagues (2021a). These datasets were chosen because they operationalized the constructs in a similar way but used a variety of scales, numbers of raters, and listener types, allowing for comparison across rating designs. Datasets 1-3, in both A and B variants, are described below in terms of the speakers, recordings, listeners, rating type, scale length, scale description, and target construct.

Dataset 1A and Dataset 1B: The first source of the data is from Authors and Colleagues (2023a), a large-scale intervention study that provided pronunciation feedback to 84 L2 students at 9 universities across the US from 17 L1 backgrounds. The goal of the study is to examine to what extent technology-based pronunciation feedback affects international teaching assistants' intelligibility improvement. A total of 15 trained raters with a background in linguistics who were graduate students or recent graduates in Applied Linguistics rated 30-second segments of the speech files for accentedness and comprehensibility, as well as completed several transcription tasks of phrases of 3-5 words in length or sentences 10-17 words in length. Raters underwent a training and scoring normalization process at the beginning of the study. The original study included both pre- and post-intervention recordings of L2 learner's unscripted classroom presentations and recorded assignments, however the present study only employs one of these recordings per participant. Roughly half (51.4%) were from the pre-intervention and the remaining were from the post intervention. Audio files were presented to the listeners in a randomized order so that the raters were not aware of the recording time when listening. The resulting recordings can be found as a part of the Authors' corpus by Authors and Colleagues

(2023b) and additional details regarding the L1 background and speaking tasks can be found in the corpus description.

Dataset 1A consists of comprehensibility ratings using a 5-point scale which ranged from 1 = *extremely incomprehensible and painstakingly effortful* to 5 = *completely comprehensible and effortless to understand* following the scale descriptors provided by Isaacs et al. (2018). However, not all raters completed this type of rating for all speakers. Of the original 84 L2 speakers and 15 raters, a subset of fully crossed rating data was extracted in which six raters listened to 39 speakers who varied greatly in proficiency level and L1 backgrounds. These six raters were 67% female and 33% male and were composed of 33% L2 English speakers with extensive experience in the North American academic context and 67% L1 English speakers with similar academic experience.

Dataset 1B is from the same study as above but contains transcriptions from a different six-rater subset of who listened to 35 recordings of sentences extracted from the first two minutes of the of the speech. This subset of raters was also 67% female and 33% male, but only one rater (17%) was and L2 speaker of English. Sentence-length stretches of speech were 10-17 words long sentences and contained minimal pauses, proper nouns, and hesitations. Listeners had only one chance to hear the recording but were able to fix misspellings before moving on to the next sample. The transcriptions were collected online using Qualtrics, and were compared to a golden transcription using a fuzzy string match approach in Python, proposed and validated by Bosker (2021). This automatic approach penalizes each character deviation, rather than word deviations, from the more established manual counts of words correctly transcribed as seen in previous research (e.g., Derwing & Munro, 1997; Munro & Derwing, 1995; Nagle & Huensch, 2020). Bosker (2021) found high correlations with human raters (r = .94), and high correlation

with acoustic phenomena required for intelligible speech when used with noisy speech recordings (See https://tokensortratio.netlify.app/ for examples and web-based processing).

Dataset 2A and Dataset 2B: The second dataset is from Authors and Colleagues (2020), a study on mobile-assisted language learning in which 16 trained raters evaluated 31 speakers for the comprehensibility and accentedness, amongst other speech perception variables. The purpose of this study was to investigate the impact of suprasegmental training on comprehensibility and intelligibility while considering learner background factors. The raters in this case were all graduate students of applied linguistics or had experience teaching or researching L2 English pronunciation. They underwent an online training and norming session prior to listening to the recordings. Thirty-seven per cent were L2 speakers of English who had lived and studied in English in North America. The original study reports on a pre and post-test, of which the present study only makes use of the pre-test data. Dataset 2A contains the comprehensibility ratings conducted by the raters using a 100-point semantic differential (i.e., 1 = *not at all comprehensible* to 100 = *very comprehensible*) digital slider scale. The scale did not have additional indications between 0 and 100, but displayed the exact value when it was manipulated by the raters. Dataset 2B comes from the same rating set but is for accentedness. The same digital slider scale approach was used but in this case the endpoints were labeled *very accented* to *not accented at all*. Additional speaker demographic information is available in Authors and Colleagues (2020).

Dataset 3A and Dataset 3B: The third dataset is part of a study of the pragmatic-speech perception relationship that was presented at the 2021 American Association of Applied Linguistics (AAAL) conference (Authors and Colleagues, 2021a). The study examined the perception of pragmatic success of responses to discourse completion tasks and relationships

between accentedness, comprehensibility, and other perceptions of the speaker with listener willingness to comply. It included 121 raters crowdsourced from Mechanical Turk (henceforth MTurk) that completed ratings of accentedness, comprehensibility, and other perceptual features of 24 L1 and L2 English speakers. The original study also included 40 undergraduate student listeners which were excluded from the present analysis in order to focus on crowdsourced listeners. It is of note that rating differences between MTurk and undergraduate students were compared, and no significant differences were found. The MTurk listeners had either no or minimal exposure to foreign accents as determined by self-report. The L2 speech samples for Dataset 3A and 3B were derived from the Authors and Colleagues (2020) experiment, however a smaller subset was included to represent a wide range of speaker proficiency and pragmatic strategies along with four L1 English speakers completing the pragmatic-focused tasks. MTurk raters also completed the ratings online but used a 9-point semantic differential scale with radio buttons instead of the sliding scale. For Dataset 3A, comprehensibility ratings were obtained using radio buttons with endpoints labeled as 1= *the speaker is easy to understand* to 9 = *the speaker is difficult to understand*. Dataset 3B contains the accentedness ratings from the study which included the same listeners and speakers included scale endpoints of 1 = *the speaker has no accent* to 9 = *the speaker has a strong accent*. See Table 1 for a summary of the datasets.

Table 1

Summary of Datasets

Dataset	Construct	N Raters	*k* Speech Samples	Scale Length	Rater type
Dataset 1A	Comprehensibility	6	39	5	Trained

Dataset 1B	Intelligibility	6	35	N/A	Trained
Dataset 2A	Comprehensibility	16	31	100	Trained
Dataset 2B	Accentedness	16	31	100	Trained
Dataset 3A	Comprehensibility	121	24	9	MTurk
Dataset 3B	Accentedness	121	24	9	MTurk

Analysis

Univariate G studies were conducted in a parallel manner on all six datasets using the *gtheory* package in R (Moore, 2016; R Core Team, 2022), and following the tutorial by Huebner and Lucht (2019). While multivariate analyses (i.e., multiple items and raters) are advantageous within the G Theory universe of analysis, this study includes only univariate G and D studies of accentedness, comprehensibility, and intelligibility ratings for comparative purposes across various rating designs with different listeners and speakers. When a univariate study is conducted, two facets are included: the rater and person (i.e., the speech sample), as well as an interaction term. The G study computes the variance explained by each facet and the interaction term in a raw variance metric and a percentage. In most L2 listener panel designs, persons (i.e., the speech samples) ideally account for more variance than do raters as raters should not be a source of inconsistent scores. The interaction term is confounded with unsystematic error and is therefore not very interpretable.

Subsequently, D studies were carried out considering Brown's (2013) recommendations and the same implementation guide in R (Huebner & Lucht, 2019). D studies predicts the changes in reliability coefficients with increasing or decreasing the number of measures within a facet. In this study, D studies were only conducted to estimate the impact of decreasing or

increasing the number of raters on generalizability and dependability coefficients. However, with a multivariate dataset and design, a study could investigate the use of multiple or different types of utterances sampled from extemporaneous speech samples. Using the process outlined by Shavelson and Webb (1991) and implemented in R by Huebner and Lucht (2019), decision tables were created for visualizable ranges of the number of raters needed to reach a target coefficient of .90. While there is much discussion about the subjectivity of setting a reliability (i.e., Cronbach's alpha) threshold in testing contexts (e.g., Brown, 2001, 2002), many L2 speech researchers have reported coefficients near .90, a target adopted by speech perception measurement in dysarthric speech measurement research (Xue et al., 2023). Additionally, both generalizability and dependability coefficients were reported because researchers may be interested in coefficients for different purposes (*generalizability* for norm-referenced tests and *dependability* for criterion-referenced tests).

Results

The current study investigated the generalizability and dependability of comprehensibility (RQ1), accentedness (RQ2), and intelligibility (RQ3) using G- and D-studies within the *Generalizability Theory* framework. The sections below are organized by research question and report first the G-Study (i.e., the components of measurement error) and then D-Study (i.e., estimations of rater panel design changes).

Comprehensibility and Rater Generalizability

For the Dataset 1A with six trained raters using a five-point scale for 39 speech samples, the results of the G study (i.e., a breakdown of measurement error components) indicate that 19.5% of variance was accounted for by the rater (r), 33.6% by the person (p), and of 46.9% by the interaction term (i.e., non-systematic variance in the data). The resulting generalizability

coefficient was acceptable at 0.81 and the dependability coefficient was 0.75. For Dataset 2A which contained 16 trained raters using a 100-point scale on 31 samples, the rater (r) accounted for 23.9% of the variance, the person (p, i.e., the speech sample) was 36.7%, and the interaction term was 39.5%. The overall generalizability coefficient was 0.94 and the dependability coefficient was 0.90. In Dataset 3A, raters accounted for even less variance than in the other two datasets. The raters (r) were 12.4% of the variance, the persons (p) were 40.7%, and the interaction term was 46.9% of the variance. The overall generalizability coefficient was .99 and dependability coefficient was also .99. See Table 2 for complete G study results. Taken together, these indicate a lack of consistency in ratings with only six raters. In order to estimate changes in the design, D studies were carried out.

The D study, which estimates impacts on coefficients if facets are changed, of Dataset 1A indicated that a substantial increase in the number of raters to about 15 to achieve the benchmark coefficient of .90. While the rubric by Isaacs et al. (2018) was not designed for high-stakes testing contexts, it is of note that even with 6 trained raters, raters are a source of nearly 20% of the variance. Dataset 2A revealed that high generalizability and dependability are achieved already at 15 raters and .80 can be achieved at about 10 raters. It appears that there is a minimal gain in generalizability beyond 20 raters. The D study results of the Dataset 3A is substantially different than the previous two. In the case of crowdsourced ratings, the generalizability and dependability coefficients do not approach .90 until about 50 raters. Beyond about 60 raters, there is very little increase in generalizability / dependability with additional raters. See Figure 1 for D study results of comprehensibility ratings of the three studies involving comprehensibility.

In sum, the results of the three comprehensibility studies indicate that a minimum of 15 listeners is necessary to achieve generalizable and dependable results (suggested to be .90 for

both coefficients, depending upon the use of the score). This number increases drastically when using crowdsourced (i.e., MTurk) listeners, requiring up to 50 to achieve the benchmark coefficient of .90. Minimal increases are estimated to be achievable with an increase in the number of listeners. Analysis across the D-study results, there is minimal divergence amongst the different scale lengths (i.e., 5-point, 9-point, 100-point), suggesting that a similar choice in the number of raters can be used for several different comprehensibility scales.

Table 2

G Study Results of Estimated Variance and Per Cent Variance Explained by Person, Rater, and Interaction Term Facets of Comprehensibility

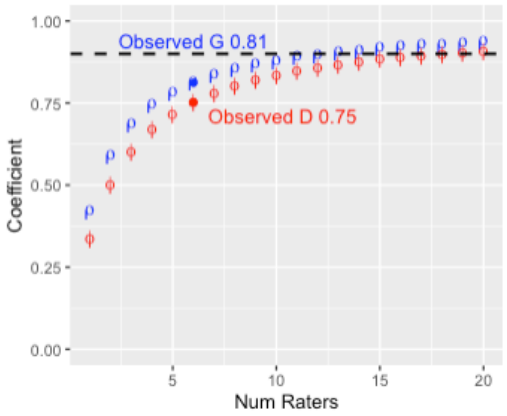
	Dataset 1A, 6 trained raters, 5-point scale, 39 samples		Dataset 2A 16 trained raters, 100-point scale, 31 samples		Dataset 3A, 121 Mturk raters, 9-point scale, 24 samples	
Effect	**Estimated Variance**	***Per cent***	**Estimated Variance**	***Per cent***	**Estimated Variance**	***Per cent***
r	0.1794421	19.5	141.2885	23.9	0.9752582	12.4
p	0.3098964	33.6	216.9071	36.7	3.2095140	40.7
rp	0.4316689	46.9	233.4750	39.5	3.6951905	46.9

Note. *r* represents rater, *p* represents person (i.e., the speech sample), and *rp* represents the interaction term

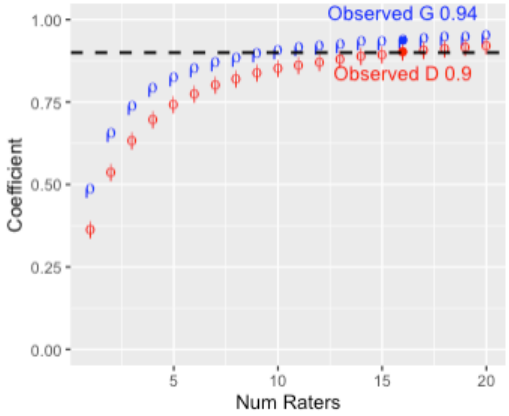
Figure 1

D Study Plots of Estimated Generalizability and Dependability Coefficients for Comprehensibility Ratings

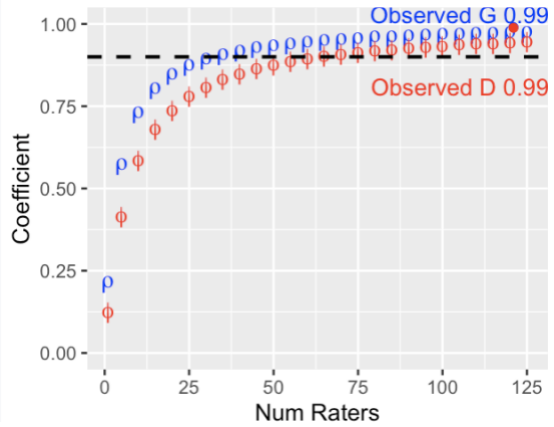
Dataset 1A, 6 trained raters, 5-point scale, 39 samples



Dataset 2A, 16 trained raters, 100-point scale, 31 samples



Dataset 3A, 121 Mturk raters, 9-point scale, 24 samples



Accentedness and Rater Generalizability

In order to investigate the related but separate constructs of accentedness, G studies were carried out on two datasets of accentedness ratings using a rater (r) by person (p) design. With 16 trained raters using a 100-point scale Dataset 2B, the facet analysis revealed that more variance was due to the raters (44.7%) than the persons (11.0%), with an additional amount to the interaction term (35.7%). While it is possible that speech file samples were limited in accentedness variation, it is not promising that more variance is proscribed to the listener than to the speaker. The overall generalizability coefficient was 0.80 and the dependability coefficient was 0.66, suggesting that accentedness rating may require significantly more raters than comprehensibility even when using trained raters.

Dataset 3B which contained ratings by 121 untrained listeners crowdsourced from MTurk reveals the impact of the increased numbers of raters. Facet analysis for this study indicated very little variance due to the rater (6.4%) and much more inter-person differences (57.9%) and a substantial interaction term (35.7%). The generalizability coefficient was 0.99 and the dependability coefficient was 0.99. While this result indicates the unnecessary use of a high number of raters, the underlying variance components can inform a D study for more efficient future rating designs. See Table 3 for person and rater facet analysis of accentedness ratings.

Table 3

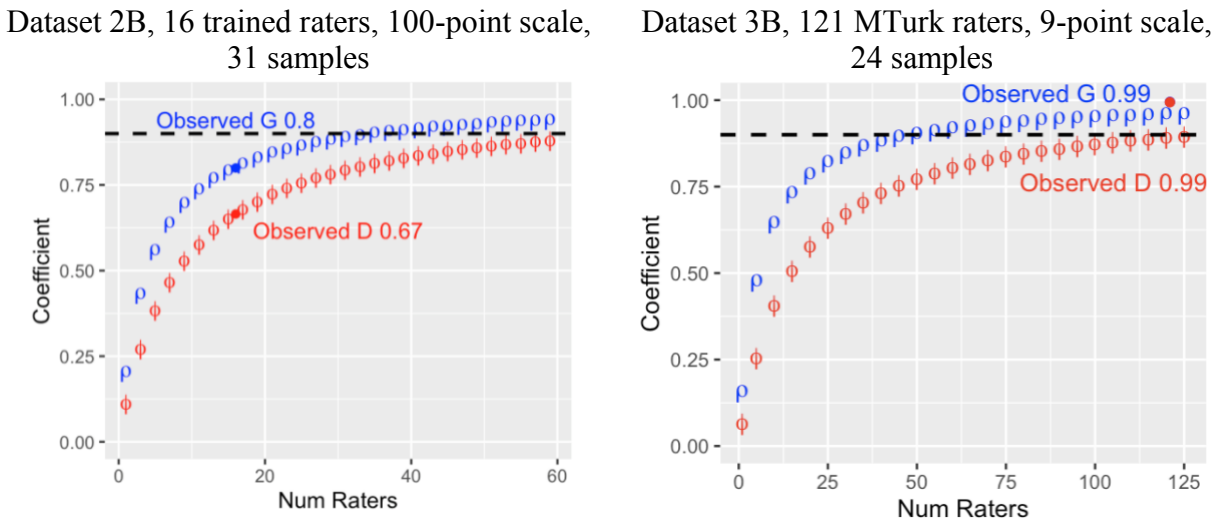
G Study Results of Estimated Variance and Per Cent Variance Explained by Person, Rater, and Interaction Term Facets of Accentedness

Effect	Dataset 2B, 16 trained raters, 100-point scale, 31 samples		Dataset 3B, 121 MTurk raters, 9-point scale, 24 samples	
	Estimated Variance	**Per cent**	**Estimated Variance**	**Per cent**
r	243.4467	44.7	0.53	6.4
p	60.2281	11.0	4.80	57.9
rp	241.3944	35.7	2.96	35.7

Note. *r* represents rater, *p* represents person (i.e., the speech sample), and *rp* represents the interaction term

Figure 2

D Study Plots of Estimated Generalizability and Dependability Coefficients for Accentedness Ratings



The D studies allow for further understanding of the impact of the number of raters evaluating accentedness. Estimates of generalizability and dependability were produced for Dataset 2B of 16 trained raters using a 100-point scale. The results indicate that about 50 raters are needed for claims of generalizability and dependability. However, the return on increasing the number of raters is minimal beyond 50. Dataset 3B which contained 121 MTurk listeners and a 9-point scale indicates that about 90 listeners are required for robust results. See Figure 2 for D study estimates of accentedness rating design adjustments.

Intelligibility and Rater Generalizability

To investigate the number of raters needed for intelligibility-based rating through transcription comparison, a G study on Dataset 1B was carried out in the same method as above. Note that this study of intelligibility has fewer listeners that the majority of L2 speech perception research, even when using trained raters. Furthermore, it uses an automated transcription scoring approach that is novel to the field (Bosker, 2021), which remains to be validated in L2 measurement. The results of the G study on Dataset 1B indicate that the rater (r) accounted for 4.0% of the variance and the person (p) (i.e., the speech sample) was 39.2%. The interaction term was 56.8% of the variance. The overall generalizability coefficient was 0.81 and the dependability coefficient was 0.79. See Table 4 for complete G study results.

Table 4

G Study Results of Estimated Variance and Per Cent Variance Explained by Person, Rater, and Interaction Term Facets of Intelligibility Transcription Tasks on Dataset 1B

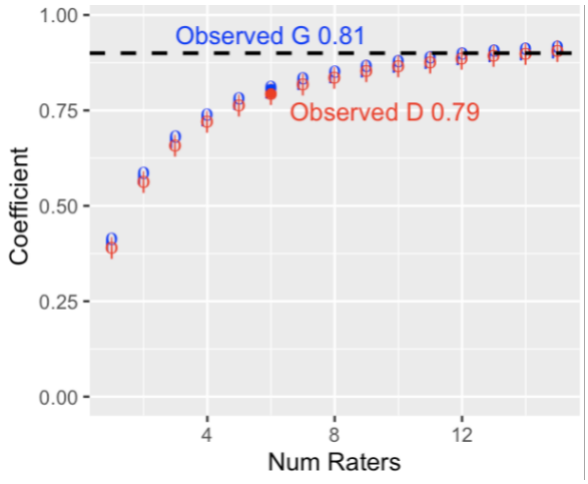
Effect	Estimated Variance	Per cent
r	3.812724	4.0
p	36.989339	39.2
rp	53.635828	56.8

Note. *r* represents rater, *p* represents person (i.e., the speech sample), and *rp* represents the interaction term

The D study results revealed that high generalizability and dependability at the .90 level is achieved with 12 listeners and .80 can be achieved at about 5 listeners. It appears that there is minimal gain in generalizability beyond 12 listeners. However, it should be noted that it is relatively common practice to extract more than one sample from a recording for listener transcription (e.g., Huensch & Nagle, 2021; Kennedy & Trofimovich, 2008; Munro & Derwing, 1995). The results from Xue et al. (2023) indicate that a second sample could enhance the dependability coefficient up to .10. See Figure 3 for scatterplot of D study estimates of Intelligibility.

Figure 3

D Study Plots of Estimated Generalizability and Dependability Coefficients for Intelligibility Transcription Tasks of Dataset 1B.



Discussion

Taken together, the G and D study results for L2 speech constructs indicate that both comprehensibility and accentedness require numerous listeners in order to reduce measurement error. The results indicate that comprehensibility may require at least 15 trained raters to achieve a generalizability coefficient of .90 when using trained raters for either the 5-point or 9-point scale. Please note that these 15 raters were experts given that they had extensive experience in L2 teaching and research, and also were trained by receiving explanations of the constructs, completing a norming session. However, when subjecting L2 speech samples to crowdsourced naïve raters, up to 50 may be required. It is of note that the use of a rubric in the five-point scale and use of a 100-point semantic differential slider resulted in the same generalizability and dependability variance structure. Unlike previous research on scale length which has found great differences in interpretation of scales (Isaacs & Thomson, 2013), these results indicate a

similarity in variance explained across these scale lengths. Table 5 summarizes the minimum number of raters required for each condition studied in this paper.

The importance of training raters becomes evident when comparing the results of the Dataset 3A with those of the Dataset 1A and 2A. In the case of untrained crowdsourced rating schemes, over 50 listeners were required. While it is possible to use crowdsourcing platforms like Mechanical Turk to gather ratings for the purposes of generalizing to larger swaths of listener backgrounds, it should be noted that this approach may introduce additional measurement noise. However, using such platforms should not be overlooked as additional steps to ensure listener attention and gather potentially important listener background variables is possible (see Nagle, 2019).

As for the second research question about accentedness, substantially more raters were needed for ratings related to the degree of accent than were for comprehensibility or intelligibility. The results indicated that 60 or more trained raters were needed, and 80 or more untrained listeners may be required for generalizable and dependable results. A possible reason is that accentedness is more intuition-based rather than a construct-based, i.e., even if each of the constructs are explained to raters with construct definitions and rating criteria, perhaps the perceived understanding of the construct itself cause the establishment of high dependability. That is, accentedness is often defined as "perception of how different a speaker's accent is from that of L1 community (Derwing & Munro, 2005)". However, this term 'being different' might be interpreted in an idiosyncratic manner for raters depending on their L1 or other individual backgrounds. Accordingly, more raters might be required for accentedness.

The construct of intelligibility, which refers to "how much the listener actually understands of the intended message" appears to require fewer listeners as it may consist of more

objectivity (Authors and Colleagues, 2018). The results indicate 12 listeners may produce consistent results sufficient enough to minimize measurement error. It is of note, however, that these findings cannot be applied to content or complete word transcription approaches for extemporaneous speech that are more common in L2 research (e.g., Huensch & Nagle, 2021; Munro & Derwing, 1995). As Kennedy and Trofimovich's (2008) transcription comparisons arrived at reliability coefficients of .72 to .84 with 12 listeners, it is possible that additional listeners may be needed to achieve a .90 target.

It is unclear with the design of this study whether the impact of training or scale length plays a larger role; however, the differences in the outcomes for comprehensibility between those with training and those without suggest that training may have a larger impact on robustness of results than does the scale length. Given the construct definition of accentedness (i.e., the strength of a foreign accent), it is somewhat unsurprising that more than 60 listeners are required to establish how strong a foreign accent might be across such a large swath of L1 English speakers. See Table 5 for a summary of the D study results.

Overall, this study is limited in a number of ways. Primarily, not all rater configurations in terms of number of items and results were able to be investigated due to the time-intensive nature of L2 speech perception research and datasets available for analysis. Furthermore, individual observations are required and as such, data sharing is limited due to IRB research protections. Additional research is necessary to confirm these findings across a wide range of rater backgrounds, speaker differences, stimulus length, and target languages. The latter point is of particular note given the globalized nature of English and the commonplace occurrences of L1 English users' experiences with accented speech. Future research could investigate the impact of

lingua franca or English medium university contexts on the generalizability of speech perception constructs.

Table 5

Summary of D study results required to reach .90 coefficients

Construct & Scale length	N trained raters	N untrained raters
Comprehensibility		
5-point	> 15	-
9-point	-	> 50
100-point	> 15	-
Accentedness		
9-point	-	> 80
100-point	> 60	-
Intelligibility		
Transcription	> 14	-

Despite these limitations, we believe that the current findings can inform the field of L2 speech in a number of ways. First, the findings of the study have provided additional evidence for validating speech constructs. There has been a long discussion about the partially independent nature of three speech constructs (i.e., comprehensibility, accentedness, and intelligibility) (Derwing & Munro, 2005). Knowing that each of these constructs requires a different number of raters to establish sufficient dependability and reliability, we can confirm that each of these constructs should be approached differently in terms of listener selection and

panel design with consideration of the listener background. Second, our findings have confirmed that rater training makes a large impact on listeners' judgements, and some constructs seem to be more consistently perceived than others. This result implies that speech ratings are a complex process which involves many multi-faceted approaches (Authors and Colleagues, 2021b). It may not be just about explaining a construct definition, but more about how raters or listeners understand the constructs based on their backgrounds. Accordingly, rater training must incorporate raters' individual and linguistic backgrounds. Furthermore, L2 speech perception researchers should be careful about generalizing their findings if they do not meet the minimum number of raters required for sufficient reliability. Finally, our results regarding listener panel designs on intelligibility should be made with consideration to the novel automated approach of scoring transcripts approach of Bosker (2021), which has yet to be validated for L2 speech.

Conclusion

This paper expanded the application of G Theory to the measurement and rating of L2 speech using six datasets of previous studies. It found that 15 or more trained listeners are required to establish robust measurement of comprehensibility when listening to one sample of a speech file. This number increases to 50 when utilizing untrained crowdsourced listeners. Furthermore, accentedness ratings require substantially more raters, 60 or 80 for trained and untrained, respectively. It also found that about 14 listeners are needed for generalizable and dependable intelligibility scores when using a transcription approach. To ensure measurements are accurate, the paper provided empirically informed recommendations for research, pedagogical, or administrative purposes.

Acknowledgements and Funding

This work is supported by NSF EAGER 2140414 funding "Second Language Speech Production: Formulation of Objective Speech Intelligibility Measures and Learner-Specific Feedback." Some of the data used for the study was also supported by the Northern Arizona University Research Bridge / SEED (RBS) Award and the Northern Arizona University Presidential Fellowship Program.

Conflicts of Interest

No, there are no conflicting interests.

References

Author (2010).

Authors and Colleagues (2023a).

Authors and Colleagues (2023b).

Authors and Colleagues (2019).

Authors and Colleagues (2018).

Authors and Colleagues (2021a).

Authors and Colleagues (2020).

Authors and Colleagues (2021b).

Bosker, H. R. (2021). Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies. *Behavior Research Methods*, *53*(5), 1945–1953. https://doi.org/10.3758/s13428-021-01542-4

Brown, J. D. (2001). Can we use Spearman-Brown prophecy formula to defend low reliability? *Shiken: JALT Testing & Evaluation SIG Newsletter*, *4*(3).

Brown, J. D. (2002). The Cronbach alpha reliability estimate. *JALT Testing & Evaluation SIG Newsletter*, *6*(1).

Brown, J. D. (2013). Score Dependability and Decision Consistency. In *The Companion to Language Assessment* (pp. 1182–1206). https://doi.org/10.1002/9781118411360.wbcla085

Brown, J. D., & Ahn, R. C. (2011). Variables that affect the dependability of L2 pragmatics tests. *Journal of Pragmatics*, *43*(1), 198–217. https://doi.org/10.1016/j.pragma.2010.07.026

Colleagues (2017). Edited by Authors.

Crick, J., & Brennan, R. (1982). GENOVA: A generalized analysis of variance system (FORTRAN IV computer program and manual). *Dorchester, MA: Computer Facilities, University of Massachusetts at Boston*.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*, *49*(4), 814–837. https://doi.org/10.1002/tesq.203

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

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

Gorsuch, G. J. (2001). Testing textbook theories and tests: The case of suprasegmentals in a pronunciation textbook. *System*, *29*(1), 119–136. https://doi.org/10.1016/S0346-251X(00)00049-X

Hubers, F., Cucchiarini, C., Strik, H., & Dijkstra, T. (2019). Normative Data of Dutch Idiomatic Expressions: Subjective Judgments You Can Bank on. *Frontiers in Psychology*, *10*, 1075. https://doi.org/10.3389/fpsyg.2019.01075

Huebner, A., & Lucht, M. (2019). Generalizability theory in R. *Practical Assessment, Research, and Evaluation*, *24*(1), 5. https://doi.org/10.7275/5065-gc10

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

Isaacs, T., & Thomson, R. I. (2013). Rater Experience, Rating Scale Length, and Judgments of L2 Pronunciation: Revisiting Research Conventions. *Language Assessment Quarterly*, *10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Isaacs, T., Trofimovich, P., & Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, *35*(2), 193–216. https://doi.org/10.1177/0265532217703433

Isbell, D. R. (2017). Assessing pronunciation for research purposes with listener-based numerical scales. In *Assessment in second language pronunciation* (pp. 89–111). Routledge.

Isbell, D. R., & Lee, J. (2022). Self-Assessment of Comprehensibility and Accentedness in Second Language Korean. *Language Learning*, *72*(3), 806–852. https://doi.org/10.1111/lang.12497

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*(3), 459–489. https://doi.org/10.3138/cmlr.64.3.459

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

Kim, Y.-H. (2009). A G-theory analysis of rater effect in ESL speaking assessment. *Applied Linguistics*, *30*(3), 435–440. https://doi.org/10.1093/applin/amp035

Levis, J. (2020). Revisiting the Intelligibility and Nativeness Principles. In *Journal of Second Language Pronunciation* (Vol. 6, Issue 3, pp. 310–328). John Benjamins. https://doi.org/10.1075/jslp.20050.lev

Miao, Y., Rose, H., & Hosseini, S. (2023). The Interaction Effect of Pronunciation and Lexicogrammar on Comprehensibility: A Case of Mandarin-Accented English. *Language and Speech*. https://doi.org/10.1177/00238309231156918

Moore, C. T. (2016). Package 'gtheory.' *Retrieved on Mar*, *15*, 2022.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech the role of speaking rate. *Studies in Second Language Acquisition*, *23*(4), 451–468. https://doi.org/10.1017/S0272263101004016

Nagle, C. (2018). Motivation, comprehensibility, and accentedness in L2 Spanish: Investigating motivation as a time-varying predictor of pronunciation development. *The Modern Language Journal*, *102*(1), 199–217. https://doi.org/10.1111/modl.12461

Nagle, C. (2019). Developing and validating a methodology for crowdsourcing L2 speech ratings in Amazon Mechanical Turk. *Journal of Second Language Pronunciation*, *5*(2), 294–323. https://doi.org/10.1075/jslp.18016.nag

Nagle, C., & Huensch, A. (2020). Expanding the scope of L2 intelligibility research: Intelligibility, comprehensibility, and accentedness in L2 Spanish. *Journal of Second Language Pronunciation*, *6*(3), 329–351. https://doi.org/10.1075/jslp.20009.nag

R Core Team. (2022). *R: A language and environment for statistical computing* (4.2.2). R Foundation for Statistical Computing. http://www.R-project.org/

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*, *n/a*(n/a). https://doi.org/10.1002/tesq.3027

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708. https://doi.org/10.1111/lang.12345

Saito, K., & Shintani, N. (2016). Foreign accentedness revisited: Canadian and Singaporean raters' perception of Japanese-accented English. *Language Awareness*, *25*(4), 305–317. https://doi.org/10.1080/09658416.2016.1229784

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using Listener Judgments to Investigate Linguistic Influences on L2 Comprehensibility and Accentedness: A Validation and Generalization Study. *Applied Linguistics*, *38*(4), 439–462. 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*. Multilingual Matters.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications.

Shin, J. (2022). Investigating and optimizing score dependability of a local ITA speaking test across language groups: A generalizability theory approach. *Language Testing*, *39*(2), 313–337. https://doi.org/10.1177/02655322211052680

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self-and other-perception of second language speech. *Bilingualism: Language and Cognition*, *19*(1), 122–140. https://doi.org/10.1017/S1366728914000832

Xi, X., & Mollaun, P. (2009). How do raters from India perform in scoring the TOEFL iBT™ speaking section and what kind of training helps? *ETS Research Report Series*, *2009*(2), i–37. https://doi.org/10.1002/j.2333-8504.2009.tb02188.x

Xue, W., van Hout, R., Cucchiarini, C., & Strik, H. (2023). Assessing speech intelligibility of pathological speech: Test types, ratings and transcription measures. *Clinical Linguistics & Phonetics*, *37*(1), 52–76. https://doi.org/10.1080/02699206.2021.2009918