Studies in Second Language Acquisition, **43** (2021), 916–939
doi:10.1017/S0272263121000292

Methods Forum

DOING L2 SPEECH RESEARCH ONLINE: WHY AND HOW TO COLLECT ONLINE RATINGS DATA

Charles L. Nagle *

Iowa State University

Ivana Rehman

Iowa State University

Abstract

Listener-based ratings have become a prominent means of defining second language (L2) users' global speaking ability. In most cases, local listeners are recruited to evaluate speech samples in person. However, in many teaching and research contexts, recruiting local listeners may not be possible or advisable. The goal of this study was to hone a reliable method of recruiting listeners to evaluate L2 speech samples online through Amazon Mechanical Turk (AMT) using a blocked rating design. Three groups of listeners were recruited: local laboratory raters and two AMT groups, one inclusive of the dialects to which L2 speakers had been exposed and another inclusive of a variety of dialects. Reliability was assessed using intraclass correlation coefficients, Rasch models, and mixed-effects models. Results indicate that online ratings can be highly reliable as long as appropriate quality control measures are adopted. The method and results can guide future work with online samples.

Since Munro and Derwing's (1995) seminal study demonstrating that intelligibility, comprehensibility, and accentedness are distinct, yet interrelated, listener-based constructs, listener-based ratings of second language (L2) speech have become one of the primary means of evaluating L2 speakers' global pronunciation skills (Saito & Plonsky, 2019). Studies have examined the linguistic features that undergird comprehensibility and

We thank Nathan Karasch and Masoud Nosrati for their help developing the Amazon Mechanical Turk interface. We also thank Tracey Derwing for her feedback on an early version of this manuscript.

The experiment in this article earned an Open Materials badge for transparent practices. The materials are available at https://osf.io/wazmc/

* Correspondence concerning this article should be addressed to Charles L. Nagle, Department of World Languages and Cultures, Iowa State University, 3102 Pearson Hall, 505 Morrill Road, Ames, Iowa 50011. E-mail: cnagle@iastate.edu

accentedness (e.g., Bergeron & Trofimovich, 2017; O'Brien, 2014; Saito et al., 2017), and listener-based ratings have been used to understand different aspects of L2 speech learning, including how L2 speakers' pronunciation develops over time (Derwing & Munro, 2013; Nagle, 2018; Saito et al., 2018) and how their impressions of one another evolve throughout a communicative interaction (Trofimovich et al., 2020).

Most studies involving listener-based ratings have recruited local listeners to evaluate L2 speech samples in a controlled, laboratory setting. On the one hand, this approach is sensible for second language (SL) contexts because it allows researchers to sample listeners from the population of individuals with whom L2 speakers are most likely to interact. On the other hand, there are many scenarios where recruiting local listeners may not be ecologically valid or even possible. Such is the case for the foreign language (FL) context, where students typically interact with one another and their instructor in a classroom setting. Practically speaking, depending on the location where the teaching and research take place, there may not be (m)any local listeners to recruit. Even if local listeners are available, they may not be familiar with the L2 varieties to which learners have been exposed, and they may not represent the individuals with whom L2 learners envision themselves interacting. Thus, there is an immediate need to devise approaches that allow researchers to locate and recruit listeners beyond the boundaries of their local community.

One promising approach is online listener recruitment through platforms such as Amazon Mechanical Turk (AMT). At least one study suggests that online ratings are reliable (Nagle, 2019), but more work is needed to understand the demographic characteristics of online listener samples, especially in L2s other than English, and how demographic characteristics and listener recruitment choices affect the reliability of the resulting data. The present study contributes to this area by comparing fully crossed ratings collected in person to ratings collected online in AMT using a pseudo-random-raters design.

BACKGROUND

LISTENER-BASED RATINGS AS A WINDOW INTO ORAL COMMUNICATIVE COMPETENCE

Listener-based ratings of fluency, comprehensibility, and accentedness have become a ubiquitous means of operationalizing L2 speakers' oral communicative competence. Fluency refers to listener's perception of the rhythm and flow of speech, comprehensibility refers to ease of understanding, or how much effort the listener has to invest to understand the speaker, and accentedness refers to the extent to which speech deviates from a target variety of the L2 (see, e.g., Derwing & Munro, 2013). These three constructs, while interrelated, capture distinct facets of oral communicative competence. For one, accented speech is often highly comprehensible (Munro & Derwing, 1995; Nagle & Huensch, 2020), and numerous studies have shown that different bundles of linguistic features predict comprehensibility and accentedness. For instance, Trofimovich and Isaacs (2012) found that rhythm (vowel reduction ratio) was the strongest predictor of accent, whereas type frequency was the strongest predictor of comprehensibility. In another study, Saito et al. (2017) reported that lexicogrammatical and pronunciation features accounted for roughly equal proportions of variance in comprehensibility (50% vs. 40%), but for accentedness, pronunciation was the primary predictor, accounting for

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

60% of variance compared with 28% for lexicogrammar. Based on these and similar results, pronunciation scholars have long recognized comfortable intelligibility, or comprehensibility, rather than accent reduction as the basic goal of pronunciation instruction (Levis, 2005).

Listener-based ratings of fluency have also played an important role in the pronunciation and speech literature. Utterance-based fluency measures (e.g., articulation rate, number of filled and unfilled pauses) are often correlated with both comprehensibility and accentedness (e.g., Saito et al., 2017; Trofimovich & Isaacs, 2012), which suggests that listeners take fluency-based measures into account when evaluating speakers along the other two scales. Put another way, fluency-based variables appear to be an important dimension of comprehensibility and accentedness. In fact, some studies have shown significant overlap in the features that predict L2 speech ratings (O'Brien, 2014), which highlights the interconnected and multidimensional nature of the three constructs.

Practically speaking, listener-based ratings of fluency, comprehensibility, and accentedness are easy to interpret and collect using simple rating scales, and the resulting data have been shown to be highly reliable. It comes as no surprise then that these constructs have had an important impact on L2 speech research and have been adapted and implemented in a range of research and teaching contexts (see, e.g., Foote & McDonough, 2017; Isaacs et al., 2017). These same reasons make listener-based ratings appealing for online research and a useful starting point for determining best practice in online data collection.

APPROACHES TO LISTENER RECRUITMENT

Recruiting an appropriate group of listeners to serve as raters means identifying the individuals with whom speakers are most likely to interact. In an SL context, the question of potential interlocutors is relatively straightforward because SL speakers routinely communicate with individuals in the local community. For instance, if SL speakers are university students, then university students can be recruited to serve as listeners (Kennedy, et al., 2015), and if they are working professionals, then listeners who work in a similar professional context can be recruited (Derwing & Munro, 2009). Recruiting an appropriate listener group is more complex in FL contexts for several reasons. For one, there may not be a single or stable target variety of the FL because FL learners have been exposed to a range of models through their instructors and may not have a clear sense of the interlocutors with whom they would like to interact in the future. Even if they do have an idea of potential future interlocutors, their imagined interlocutor group is likely to change as their learning goals evolve. In light of these considerations, various approaches to listener recruitment are possible. Recruitment could be guided by the L2 varieties to which FL learners have been exposed or by the types of interlocutors with whom they envision themselves interacting. Another alternative would be to recruit a diverse listener group because FL learners may end up interacting with many different types of interlocutors as they become more proficient L2 users.

It bears mentioning here that listener recruitment choices have implications for construct definitions. For example, Derwing and Munro defined accentedness as "how different the speakers' accents are from a standard Canadian English accent" (2013, p. 185). However, for FL research, a standard local variety may not be a sensible anchor

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

point because learners' speech patterns are likely an amalgamation of the diverse dialects to which they have been exposed and thus may not be aligned with any single native variety of the L2. In fact, eliciting accentedness ratings relative to a local standard may underestimate participants' pronunciation ability if, for example, a listener perceives a relatively nativelike accent as moderately to strongly accented because it does not coincide with the local norms with which that individual is familiar. Instead of accentedness relative to a local standard, the notion of foreign accent, or the degree to which the speaker's accent deviates from any native variety of the L2, may be suitable for FL studies. However, foreign accent ratings are not without their pitfalls. When provided by a diverse group of listeners, such ratings may be noisy because it is unlikely that all listeners would be equally familiar with the speech characteristics of other L2 dialects. There is also the issue of precisely how to define foreign accent so that listeners understand foreign to mean nonnative rather than a native speaker who is not from the local area.

Once these methodological and operational issues are resolved, there is the practical task of actually locating and recruiting listeners. In SL contexts, this means turning to the local community. In FL contexts, listener recruitment can be difficult. In many locations, there may not be any native listeners to recruit, and even if native listeners are available, they may not match the target listener profile. FL researchers could travel to a location where the L2 is spoken or rely on their colleagues abroad, but those options are not time- and cost-effective. Recruiting listeners online is a practical alternative.

APPROACHES TO THE RATING DESIGN AND PROCEDURE

Although scalar ratings are relatively easy to implement, their apparent simplicity belies several complex decisions that must be made. How many points should the rating scales include? Should ratings be carried out simultaneously or sequentially? And should the rating design be fully crossed, such that all listeners evaluate all items, or can a random-raters design be used? Fortunately, researchers have begun to address these questions. With respect to scale length, Munro (2017) found that 18 of 21 listeners used at least nine choices in rating comprehensibility, suggesting that a 9-point scale would be the minimum number of steps needed for sufficient resolution (see also Southwood & Flege, 1999). However, in another study, Isaacs and Thomson (2013) found that a 9-point scale resulted in fuzzier distinctions between steps than a 5-point scale, which they argued could have been due to the relative homogeneity of the speaker sample (i.e., the speakers did not show a wide enough proficiency range to warrant nine distinct options). Other researchers have used 1,000-point slides and obtained results that fall in line with studies using shorter interval scales. In sum, then, appropriate scale length depends on other study features (e.g., the anticipated proficiency spread of the speakers), although for most research a scale of at least 9 points seems advisable. Regarding the rating procedure, in a study on sequential versus simultaneous ratings, O'Brien (2016) found that the two approaches yielded comparable results.

The third question on the ratings design has not been addressed in the literature. In a fully crossed design all listeners evaluate all items, whereas in a random-raters design a random subset of listeners evaluates each item, such that all items are evaluated many times, but the listener groups that rate each item are different. Fully crossed designs, which are common in L2 research, are advantageous because they allow for rater-by-item effects

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

to be taken into account during data analysis. However, there are many instances in which fully crossed designs may not be feasible, such as studies that generate a large number of files to be evaluated. In that case, raters could adopt a blocked design by randomizing files into blocks to be evaluated by different listener groups. Typically, each group rates a subset of common files shared across blocks, allowing for robust estimation of reliability, as well as a set of unique files available only to that group (see, e.g., Trofimovich et al., 2009; Wisniewska & Mora, 2020). In a completely random-raters design, each item would be evaluated by a random group of *k* raters, such that no two items share the same rater group. Thus, there are a range of options that researchers can leverage depending on their needs.

EXPANDING THE TOOLKIT: ONLINE APPROACHES

Online platforms such as AMT offer researchers several practical and methodological advantages. For one, they can connect researchers with a large pool of potential raters to whom they might not otherwise have access, which may be especially important for FL researchers. They are also readily scalable to the size of the study, insofar as researchers can collect a large number of ratings relatively quickly. At the same time, in an online design, researchers cannot directly oversee data collection and thus have limited insight into how listeners carry out the ratings. Therefore, appropriate quality control measures must be put into place.

One common quality control measure is an instructional manipulation or attention check. This measure involves inserting directions on how to respond to the item into the item itself. Raters who follow the directions are classified as attentive, whereas raters who do not are classified as inattentive. The data from the latter group are then removed before analysis. Studies have shown that laboratory and online participants perform similarly on simple attention checks (Goodman et al., 2013; Paolacci et al., 2010). However, such post-hoc screening strategies can result in substantial data loss because data provided by inattentive workers must be excluded from analysis. Post-hoc strategies have also been criticized because they may fundamentally alter the demographic characteristics of the listener sample, which can have an impact on findings (Paolacci & Chandler, 2014). For these reasons, researchers have advocated for pre-screening measures, such as making tasks available to online workers who have been vetted. For instance, AMT allows requesters to limit tasks to high-reputation workers whose overall approval rate for work completed exceeds a certain threshold (e.g., 90%). Data provided by high-reputation workers have shown to be highly reliable, eliminating the need for attention checks (Peer et al., 2014). The challenge with recruiting only high-reputation workers is that such workers may not be available in all L2s because a large portion of the AMT userbase consists of English-speaking individuals (Ross et al., 2010). Thus, other pre-screening methods may need to be developed for studies that aim to recruit non–English-speaking workers.

Researchers who work in AMT and other online platforms must also consider the ethical dimension of their approach. Historically, AMT workers have been characterized as hobbyists, individuals who participate in AMT for fun rather than income. In fact, this view has been debunked. The reality is that many AMT workers do rely on the wages they receive (Martin et al., 2014). Moreover, as Fort et al. (2011) pointed out in their critical

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

assessment, AMT workers lack standard workplace protections. For example, when creating a task, requesters can create a review period during which time they can review work submitted and decide to approve it or reject it without pay, but workers have little recourse to address concerns that they have about requesters. Researchers can mitigate some of these concerns by paying workers a fair wage commensurate with the complexity of the task (e.g., at least the federally mandated minimum wage, which in the United States is $7.25 per hour at the time of writing) and paying them promptly for all work completed, even if the work does not appear to be completed correctly. The latter is particularly important because incorrectly completed assignments could be due to several issues beyond the worker's control, such as the instructions that the researcher provided or the interface itself. This is why it is also useful to include an open-ended response box where workers can provide feedback and recommendations to guide future improvements.

ONLINE SPEECH RESEARCH

A growing body of work has explored the utility of AMT for linguistic research (e.g., Callison-Burch & Dredze, 2010). L2 speech researchers have used AMT to identify and grade mispronunciations, with the goal of using the crowdsourced data to improve computer-assisted pronunciation training systems (Peabody, 2011; Wang et al., 2013). Researchers have also crowdsourced intelligibility data using transcription tasks as well as accuracy and comprehensibility data using a scalar ratings interface (Loukina et al., 2015; Nagle & Huensch, 2020). Although researchers are increasingly turning to AMT and other online platforms, to date, only one study has reported on the steps needed to design the AMT interface, collect and process the data, and examine its reliability.

In Nagle (2019), 50 speech samples (39 L2 samples, 4 near-native anchor samples, and 7 attention checks) were paired with a simple AMT rating interface where workers were allowed to listen to the sample up to three times before evaluating it for comprehensibility, fluency, and accentedness using separate 9-point scales. A completely random-raters design was adopted, where each file was evaluated by a unique group of 20 workers, and the task was made available only to AMT workers whose Internet protocol (IP) address was located in a country where Spanish was an official language. Of 54 AMT workers, only 4 were classified as inattentive, but an additional 15 had to be excluded because they did not complete the minimum number of attention checks required to evaluate the quality of their work (12 workers) or because they did not rate the minimum number of near-native anchor clips required to determine that they had understood the instructions and rating scales (3 workers).

Intraclass correlation coefficients (ICC) were used to estimate reliability, and Rasch models were fit to the data for each construct for the 35 workers who were retained after implementing the quality control measures. Results indicated excellent reliability for all three constructs (in all cases, ICC > .87), but Rasch modeling revealed some issues with scale use and structure. Namely, the 9-point scales did not yield sufficiently distinct steps, especially for accentedness. Based on these results, Nagle (2019) made several recommendations to improve data retention and validation in AMT, including creating a screening task to award workers a special qualification necessary to advance to the rating experiment and blocking files to avoid excluding workers who evaluated a small number of samples (e.g., <3).

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

THE CURRENT STUDY

Building upon Nagle (2019), the present study was borne out of the practical need to continue to refine a robust approach to conduct L2 speech research online. Developing a method for online data collection offers researchers a complementary tool that they may choose to use under certain circumstances (e.g., if they cannot recruit an appropriate group of raters locally, if they have collected a large number of samples to be evaluated). We implemented Nagle's (2019) suggestions by (1) creating a screening task where workers completed a comprehensive background survey and rated a small number of samples to familiarize themselves with the instructions and interface and (2) blocking files to ensure that all workers evaluated a similar number of files (i.e., leading to a pseudo-random-raters design). We also made improvements to the AMT interface, directly incorporating certain quality control measures (e.g., timers to ensure that raters moved through the task at a reasonable pace). We collected data from two AMT listener groups, a learner-guided dialect group composed of AMT workers representing the dialects to which learners had been exposed (i.e., the dialects that their instructors spoke) and an any-dialect group composed of AMT workers recruited from all Spanish-speaking countries. We also collected data from a group of local laboratory listeners for comparison. These listeners were necessarily drawn from many different Spanish dialects because it would have been difficult to recruit a homogenous and ecologically valid rater group at the location where the research took place. This study was, therefore, guided by the following research questions:

1. What are the demographic characteristics of AMT workers based in Spanish-speaking countries?
2. How reliable are the comprehensibility, fluency, and foreign accent ratings provided by each group (lab, AMT learner-guided dialects, and AMT any dialect of Spanish)?
3. What is the minimum number of raters needed to be recruited online to establish a reliable aggregate rating?

To answer the first research question, we descriptively analyzed the background characteristics of participating AMT workers. To answer the second research question, we conducted three separate analyses. First, we computed standard reliability coefficients for each listener group. Then, we fit separate Rasch models to each group to examine differences in rater severity and fit. The Rasch models provided an additional perspective on scale fit and reliability for each construct. Finally, we fit a linear mixed-effects model and carried out post-hoc comparisons to determine if there were significant between-group differences in the way listeners scored the speakers on each construct. To answer the third research question, we resampled our data at progressively smaller listener sample sizes (e.g., n = 20, 19, 18), recalculating reliability at each step.

METHOD

SPEECH SAMPLES

Twenty-three speakers who were recruited from multiple sections of two intermediate-level Spanish language courses provided the speech samples used in this study. After watching a silent animated short film about a girl who had to defeat a monster that sprang

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

out of her journal, speakers received a set of eight screenshots captured from the short and used them to retell the story.[1] Speakers were given five keywords to help them retell the story and had up to a minute to look over the screenshots before they were recorded. During the planning time, they were not allowed to take notes. Speakers were recorded using a high-quality, head-mounted microphone connected to a desktop computer in a sound-attenuated room. Following standard procedures, we prepared the audio files for listener evaluation by creating a 30-s sample from the beginning of each participant's full recording, excluding initial pauses and hesitations. We then normalized all samples to a comfortable listening volume. We used one of the 23 samples as a practice file and divided the remaining 22 samples into two blocks of 11 L2 audio files to be used in the AMT experiment. Three native speakers of Argentinian Spanish provided the control samples. We included samples from two of the speakers in the experimental blocks, reserving the sample from the third native speaker as a practice file.

LISTENERS

Three listener groups participated in this study: lab listeners (Lab); AMT listeners who were recruited from Spain, Argentina, and Mexico, the Spanish dialect regions to which learners had been exposed (AMT L-Guided); and AMT listeners representing a range of dialects (AMT Any Dialect).[2] We provide a detailed description of how we recruited and screened online listeners in the procedures section. Here, we focus on the characteristics of the listeners who participated in the experimental portion of the study after passing the screening task.

The Lab listeners were 14 native Spanish speakers who were pursuing an advanced degree at the university where the research took place. They reported the following countries of origin: Colombia (4), United States (3), Peru (2), Spain (2), Ecuador (1), Costa Rica (1), and Mexico (1). Although we recruited and approved AMT L-Guided listeners from Spain, Argentina, and Mexico, of the 25 listeners who completed the experimental rating task, 22 were from Spain, 2 were from Mexico, and 1 was born in Venezuela but was residing in Spain at the time. The fact that most listeners in this group were from Spain, and none were from Argentina, highlights one of the challenges of recruiting balanced listener groups (at least with respect to dialect) online, a point we return to in the discussion. Finally, the 23 AMT Any Dialect listeners who completed the experimental rating were from the following regions: Spain (11), Colombia (5), Chile (2), Mexico (2), Argentina (1), Ecuador (1), and Venezuela (1).

The demographic characteristics of the listener groups are summarized in Table 1. All three groups reported exposure to English between the ages of 5–7 and rated their English speaking and listening skills in the upper range ($M > 6.00$) on the 9-point proficiency scale (*extremely low proficiency–extremely high proficiency*). As expected, the Lab listeners, who were living in the United States and pursuing a graduate degree at a US university, evaluated their English skills slightly more positively than the AMT listeners did. The Lab listeners reported interacting in English far more frequently than in Spanish, whereas the opposite was true for the AMT listeners. In fact, patterns of language use (percent daily use of English and Spanish) for the Lab and AMT listeners were near mirror images of one another. The Lab listeners also reported less familiarity with L2 Spanish speech, which is likely due to the fact that they spent most of their time interacting in English and would not

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

TABLE 1. Listeners demographics

	AMT L-Guided (n = 25)		AMT Any Dialect (n = 23)		Lab (n = 14)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Biological age	33.92	9.86	32.74	8.81	28.71	6.74
Age onset: English	5.68	3.34	7.22	3.48	5.14	5.30
English listening	6.68	1.44	6.78	1.78	7.71	1.20
English speaking	7.36	1.25	7.26	1.66	7.57	1.40
% Daily Spanish	78.12	20.32	79.35	17.01	22.00	11.12
% Daily English	17.24	16.87	15.87	14.82	77.86	11.22
% Daily other	5.04	16.57	4.78	11.92	0.14	0.53
Familiarity L2 speech	6.44	2.10	6.70	2.01	5.36	2.17
Frequency L2 interaction						
Never	3		2		2	
1/Month	13		13		9	
1/Day	4		2		3	
>1/Day	5		6		0	
Context L2 interaction						
NA	3		2		2	
Personal	9		9		7	
Professional	7		8		3	
Both	6		4		2	
Linguistics	Yes: 8	No: 17	Yes: 10	No: 12	Yes: 7	No: 7
Language teaching	Yes: 4	No: 21	Yes: 2	No: 21	Yes: 8	No: 6

have had much opportunity to interact with non-native Spanish speakers in Spanish in an English-dominant environment.

Regarding context of interaction, there was a trend toward the personal domain for the Lab listeners, whereas the AMT listeners showed a relatively even spread across the categories, albeit with slightly fewer workers reporting interacting with non-native speakers in both personal and professional contexts. In all three groups, there were very few individuals who reported no interactions with non-native speakers. Last, a third to half of the listeners in the three groups reported linguistic experience, but teaching experience was more common for the Lab listeners than for the AMT workers.

RATING TASK

All materials associated with the AMT rating interface, including the html code to generate the tasks (Spanish and English versions), a document outlining the task properties that were implemented when the tasks were deployed in AMT, and a guide for modifying the interface are available in the Online Supplementary Materials. Study materials can also be accessed at https://www.iris-database.org and https://osf.io/wazmc.

Working with a computer programmer, we designed an AMT rating interface consisting of the following elements: (1) an informed consent document; (2) a comprehensive background survey; (3) an overview of the speaking task, including an embedded copy of the animated short and the screenshots speakers received, a summary of the constructs to be evaluated with instructions and information on the rating interface; (4) two practice files to be evaluated, one from an L2 speaker and one from a native speaker, neither of

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

whom provided files for the experimental rating task; (5) the experimental files; and (6) a posttask survey. We adapted Derwing and Munro's (2013) construct definitions. We adhered to their definitions for fluency and comprehensibility, but we modified the accentedness scale to target foreign accent, which we defined as any pronunciation feature that would not occur in native Spanish speech. We also instructed listeners that assigning the audio file the best possible score on the foreign accent scale would signify that the speaker could be a native speaker of Spanish. In this way, we aimed to sensitize listeners to the distinction between pronunciation features that would indicate a nonnative speaker versus pronunciation features that could correspond to a native variety of Spanish. We gave listeners the following definitions (backtranslated from Spanish; for the Spanish version, see the HTML task preview):

- Fluency: Fluency refers to the rhythm of the language, that is, if the speaker expresses themselves with ease, or if they have difficulty expressing themselves and pause often.
- Comprehensibility: Comprehensibility refers to how easy or difficult it is to understand what the speaker is saying. You may be able to understand everything the speaker says, but doing so required a lot of attention and effort on your part. What we are interested in is how much effort you have to expend to understand the speaker.
- Foreign Accent: We all have an accent, but for our purposes, we are interested in foreign accent, that is, any pronunciation feature that does not occur in the speech of a native Spanish speaker. Keep in mind that foreign accent is different from comprehensibility: it could be that the speaker is easy to understand even if they speak with a strong foreign accent.

Each audio file was presented individually on a rating screen with a play button and 7-point fluency, comprehensibility, and foreign accent scales arranged horizontally. We selected 7-point rating scales to strike a balance between a lower-resolution 5-point scale, which may not have given listeners enough options, and a higher-resolution 9-point scale whose steps may have overlapped for the intermediate speakers who provided our speech samples (Isaacs & Thomson, 2013). Anchors were provided only at the extremes, where higher scores were always better on the 7-point scales: for fluency, *not very fluent–very fluent*; for comprehensibility, *very difficult to understand–very easy to understand*; and for foreign accent, *very strong foreign accent–no foreign accent (could be a native speaker of Spanish).* The instructions that appeared on the rating screen made it clear that the scales would only become active after the audio had finished playing.

After the file played through, workers had up to 45 s to make their ratings before the page became inactive. These two timers were quality control measures that served a similar function to Nagle's (2019) attention control checks, insofar as they ensured that workers listened to the entire audio file before rating it and moved through the task at a reasonable pace. All experimental audio files were presented in a unique random order to each worker. After completing the ratings, workers received the posttask survey where they rated their understanding of the comprehensibility and foreign accent constructs and the difficulty the task posed using 100-point sliders and optionally provided open-ended feedback on any aspect of the interface or procedure. It was not possible for the in-person raters to complete the AMT version of the rating task. We, therefore, developed a Qualtrics survey whose format mirrored that of the AMT interface.

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

PROCEDURE

We recruited local listeners from the same large, public university where we recruited the speakers. We sent an email to 3,519 graduate students with information about the study and posted study information to relevant university message boards. We also relied on word of mouth and our professional networks. We were ultimately able to recruit 14 local listeners who were native speakers of Spanish. The second author met with the Lab listeners individually in a quiet space for data collection.

As shown in Figure 1, we split the AMT interface into two tasks. The first task was a screening measure consisting of the informed consent document, background survey, instructions and information on the rating interface, and two practice files (one from an L2 speaker and the other from a native speaker). To recruit AMT L-Guided listeners, we deployed the task to 100 workers located in Argentina, Mexico, and Spain using AMT geographic filters. The task was active for 7 days before the 100-worker completion threshold was met. We used the screening data to validate workers, assigning them a study-specific qualification that allowed them to view and complete the experimental ratings task. Workers had to meet three criteria to receive the study-specific qualification: (1) Spanish had to be their native language (or one of their native languages), (2) they had to be born in a Spanish-speaking country, and (3) they had to rate the native speaker practice file better than the learner file on all three scales, which we took as an indicator that they had understood the instructions and used the scales properly (i.e., they did not reverse the directionality of the scales). Applying these criteria, we eliminated five workers whose L1 was not Spanish, six L1 Spanish workers who were not born in a Spanish-speaking country, and 11 L1 Spanish workers who assigned the learner practice file a higher score than the native speaker practice file. Regarding the latter, all 11 cases were related to the foreign accent scale, where workers reversed the scale, interpreting lower values as indicative of a better score (i.e., less foreign accent) than higher values, despite instructions to the contrary. Thus, 78 AMT L-Guided workers were approved to advance to the experimental task.

The experimental task consisted of an informed consent document, the same instructions and information that workers had received on the screening task, the experimental files to be rated, and a post-task survey. Because we blocked the 24 samples into two

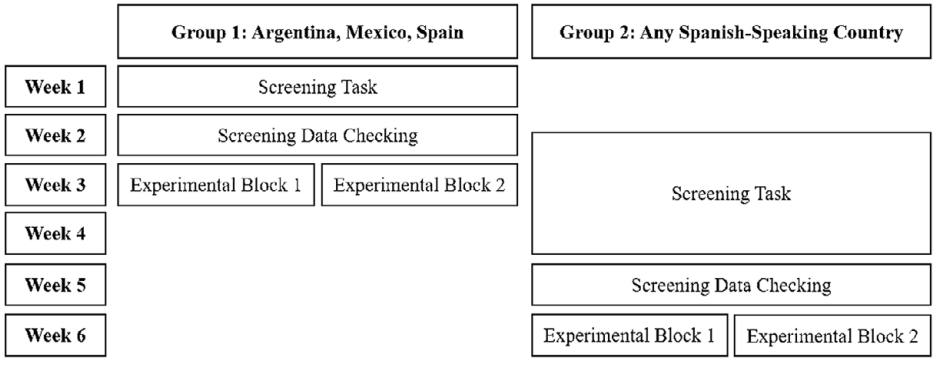

FIGURE 1. Overview of the structure and timing of the online ratings in Amazon Mechanical Turk.

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

groups of 12 files (11 L2 files and one native file), we deployed two versions of the experimental task, one per block, simultaneously, requesting 20 workers per block. Some workers completed both versions, rating all 24 samples. Of the 78 screened and approved AMT L-Guided workers, 25 completed the experimental ratings task.

Following the same procedure, we created the AMT Any Dialect group by recruiting and validating AMT workers located in any Spanish-speaking country. We staggered data collection for the AMT L-Guided and AMT Any Dialect groups to prevent workers from participating in the experiment twice. Thus, after we had recruited workers for the AMT L-Guided group, we awarded all participating AMT L-Guided workers a special qualification to prevent them from retaking the screening task. We then deployed the screening task a second time to 100 workers located in any Spanish-speaking country. The second screening task was active for 17 days before 100 new workers were recruited. Of those 100 individuals, 6 were excluded because their L1 was not Spanish, 8 because they were not born in a Spanish-speaking country, and 24 because they did not score the native speaker practice file higher than the learner file on all three rated dimensions. Thus, 62 AMT Any Dialect workers were approved to advance to the experimental task. Of those 62 workers, 23 completed the experimental task.

Because the experimental rating task involved a small number of speech samples, both Lab and AMT listeners evaluated all files in a single sitting without a break. The experiment was self-paced, insofar as listeners could move from one file to the next at a pace that felt comfortable, but the experimental rating session was time-controlled. AMT workers had up to 30 min to complete the experimental rating task before the task became inactive (AMT allows researchers to specify a time within which the task must be completed), and Lab listeners were kept on pace by a research assistant who supervised the experimental session. Lab listeners wore noise-canceling headphones while completing the task, and AMT workers were instructed to wear headphones while carrying out the ratings.[3] AMT workers were compensated at a rate of $7.25 per hour following the US federal minimum wage at the time of recruitment, and Lab listeners received a $10 honorarium.

RESULTS

WORKER DEMOGRAPHICS: SCREENED AND APPROVED AMT WORKERS

Because most AMT research has recruited native English speakers, we were interested in examining the demographic characteristics of AMT workers who passed the screening task (i.e., AMT workers whose L1 was Spanish, who were born in a Spanish-speaking country, and who correctly used the rating scales). As shown in Table 1, the two AMT groups began learning English relatively early in life (in both cases, $M_{ageonset} < 7$ years). Means for self-estimated listening and speaking proficiency in English exceeded 6.00 for both groups on the 9-point scale (anchors: *extremely low proficiency–extremely high proficiency*). As expected, the AMT workers, all of whom were located in a Spanish-speaking country, reported using mostly Spanish in their daily interactions, followed by English, and some additional languages (e.g., Catalan, Galician, and Basque, the three regional languages spoken in Spain; Portuguese; German; and other L2s that they had learned).

Both groups reported approximately the same degree of familiarity with L2 Spanish speech ($M = 6.63$ and 6.74 for the L-Guided and Any Dialect groups, respectively). The

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

TABLE 2. Listener demographics: screened and approved AMT workers

	AMT L-Guided (n = 78)		AMT Any Dialect (n = 62)	
	M	*SD*	*M*	*SD*
Biological age	32.59	8.62	32.77	8.52
Age onset: English	6.41	4.13	6.84	3.81
English listening	7.31	1.39	7.18	1.70
English speaking	6.72	1.63	6.63	1.79
% Daily Spanish	80.82	17.70	76.06	19.33
% Daily English	15.05	13.80	17.40	16.45
% Daily other	4.90	12.57	8.79	18.27
Familiarity L2 speech	6.63	1.80	6.74	1.93
Frequency L2 interaction				
Never	8[1]		7	
1/Month	36		34	
1/Day	25		9	
>1/Day	9		12	
Context L2 interaction				
NA	6[1]		7	
Personal	22		20	
Professional	27		15	
Both	23		20	
Linguistics	Yes: 37	No: 41	Yes: 29	No: 33
Language teaching	Yes: 9	No: 69	Yes: 9	No: 53

Note. [1] Two workers who indicated that they never interacted with L2 speakers on the frequency of interaction item nevertheless reported both personal and professional interactions with non-native speakers on the context of interaction item.

most common frequency of interaction with non-native speakers was once per month, but a sizable portion of workers in each group reported daily interaction. For the L-Guided group, this amounted to nearly half of workers compared with approximately a third of the Any Dialect workers. Context of interaction was largely balanced across the three categories. Most AMT workers had some background in linguistics (i.e., they had taken a course that dealt with linguistic topics), but few reported language teaching experience. Overall, then, the characteristics of the online AMT workers who passed the screening task were in line with the characteristics of the subset of workers who completed the experimental task (cf. Table 1).

DESCRIPTIVE STATISTICS: L2 SPEECH RATINGS

As a first step toward validating the data, we computed descriptive statistics for each group. As reported in Table 3, the two native speaker files received much higher ratings on average than the learner files, and, as shown in Figure 2, the modal response for the native speaker files was 7, the highest possible score, for all groups on all constructs. Figure 2 also shows that the overall distribution of scores for each construct was similar across the three listener groups: Comprehensibility scores were relatively distributed throughout the 7-point continuum, and fluency and foreign accent scores were slightly and strongly skewed toward the less fluent/stronger foreign accent end of the continuum.

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

TABLE 3. Means and (*SD*s) by rater group and construct

	Comprehensibility		Fluency		Foreign accent	
	L2	NS	L2	NS	L2	NS
AMT L-Guided	3.40 (1.28)	6.62 (0.91)	3.10 (1.25)	6.41 (1.00)	2.42 (1.18)	6.44 (0.89)
AMT Any Dialect	3.40 (1.41)	6.58 (0.84)	2.93 (1.20)	6.70 (0.52)	2.19 (1.07)	6.48 (0.78)
Lab	3.63 (1.62)	6.96 (0.19)	3.28 (1.42)	6.82 (0.48)	2.41 (1.28)	6.79 (0.50)

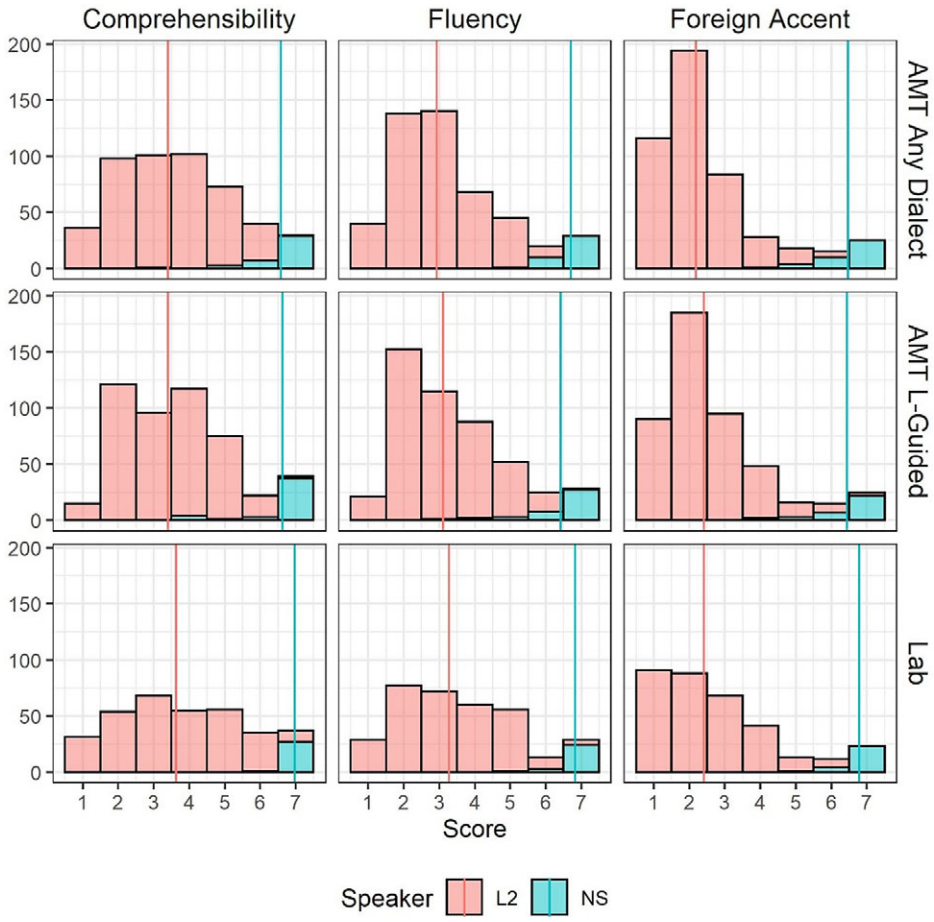

FIGURE 2. Score distribution by rater group and construct.

RELIABILITY COEFFICIENTS

To examine reliability, we computed two coefficients: the two-way, consistency, single-measure intraclass correlation coefficient (ICC(C, 1)), and the two-way, consistency, average-measure intraclass correlation coefficient (ICC(C, *k*)). The single-measure ICC is

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

TABLE 4. Reliability coefficients by rater group and construct

	Comprehensibility		Fluency		Foreign Accent	
	(C, 1)	(C, *k*)	(C, 1)	(C, *k*)	(C, 1)	(C, *k*)
AMT L-Guided	.35	.92	.38	.92	.18	.81
AMT Any Dialect	.31	.90	.35	.92	.14	.76
Lab	.56	.95	.48	.93	.35	.88

Note. (C, 1): Two-way, consistency, single-measure intraclass correlation coefficient. (C, *k*): Two-way, consistency, average-measure intraclass correlation coefficient.

an estimate of the reliability of ratings provided by a single individual, and the average-measure ICC is an estimate of the reliability of ratings provided by a group of *k* individuals. Cicchetti (1994) proposed the following cutoffs for the ICC: <.40 = poor, .40–.59 = fair, .60–.75 = good, and >.75 = excellent. As shown in Table 4, average-measure coefficients were all in the excellent range, whereas most of the single-measure coefficients were in the poor range. With respect to the online groups, the reliability coefficients for the AMT L-Guided group were slightly higher than the coefficients for the AMT Any Dialect group.

RASCH MODELS

We fit three separate Rasch models to examine rater severity, fit indices, and scale use for each of the three listener groups. Each model included three facets: examinees (i.e., speakers), raters (i.e., listeners), and scale categories (i.e., comprehensibility, fluency, and foreign accent). A fixed chi-squared test of the null hypothesis that the Lab listeners were of the same severity level was significant ($\chi^2(13) = 501.5$; $p < .001$). In other words, the Lab listeners showed statistically different levels of severity. The logit measures associated with the rater facet provides information on individual listeners and their respective severity levels. The overall range was 3.21 logits, from –0.76 for the most lenient Lab listener to 2.45 for the most severe Lab listener. The separation index, which shows the number of severity levels, was 6.07 with a reliability estimate of .97, suggesting that there were approximately six statistically distinct levels of severity. As for rater fit statistics, a range of 0.50–1.50 for infit values can be interpreted as good internal consistency (Eckes, 2015), and all but one listener (infit = 2.08) was in that range. With respect to the rating categories, foreign accent was the most severely rated (logit value of 0.78), which is in line with previous research (e.g., Munro & Derwing, 1995; Nagle & Huensch, 2020), while comprehensibility and fluency yielded more lenient ratings (logit values of –0.57 and –0.21, respectively). According to Eckes (2015), rating scale effectiveness may also be examined through fit statistics such as the mean-square outfit statistic, which should not exceed 2. For each of the three rating categories, the rating scale had an excellent model fit; values of the outfit mean-square statistic were 1.33, 0.95, and 0.77 for foreign accent, comprehensibility, and fluency, respectively.

A fixed chi-squared test of the null hypothesis that AMT Any Dialect listeners were of the same severity level was significant ($\chi^2(22) = 533.6$; $p < .001$). This means that, like the Lab listeners, the AMT Any Dialect listeners showed different levels of severity. The logit

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

measure range (4.60) was larger than the range for the Lab listeners, with a value of 0.34 for the most lenient AMT Any Dialect listener and a value of 4.26 for the most severe listener. Such a high upper logit indicates that this group of AMT listeners may have been more severe than the Lab listeners. Despite the larger logit range, the separation index of 5.48, with a reliability estimate of .97, indicates that there were five to six statistically distinct levels of severity. With respect to rater fit statistics, the mean rater infit value of 1.07 points to good internal consistency, but four AMT Any Dialect listeners were just outside the suggested range, with infit values of 0.45, 0.46, 1.67, and 1.70. As was the case for the Lab listeners, this group of AMT listeners was the most severe when rating foreign accent (logit value of 0.82), followed by fluency (–0.16) and comprehensibility (–0.67). Outfit mean-square statistics of 1.13, 1.07, and 0.90 were obtained for the foreign accent, comprehensibility, and fluency scales, respectively, indicating excellent scale fit.

The AMT L-Guided listeners also showed statistically different levels of severity ($\chi^2(23) = 391$; $p < .001$). Their logit measure range was 2.18, from –0.15 for the most lenient to 2.03 for the harshest rater in this group. The separation index of 3.75 with a reliability of 0.93 suggests that there were approximately four different levels of severity. Additionally, four AMT L-Guided listeners exhibited overfit, with infit values slightly under 0.50 (0.26, 0.34, 0.42, and 0.41). Category severity showed a similar pattern to the other two listener groups: foreign accent (logit value of 0.62), fluency (logit value of –0.17), and comprehensibility (logit value of –0.46). Outfit values for rating categories also pointed to an excellent model fit, with values of 1.30, 0.98, and 0.80 for foreign accent, comprehensibility, and fluency, respectively.

To sum up, although distinct levels of severity were observed within each listener group, the Rasch models suggested good overall performance for raters and scales. Whereas approximately six distinct levels of severity were observed for the Lab and AMT Any Dialect listeners, the AMT L-Guided group showed only four severity levels, suggesting greater uniformity in their ratings.

MIXED-EFFECTS MODELING

To determine if the three listener groups rated the L2 speakers differently on each construct, we fit a linear mixed-effects model in R version 4.0.2 (R Core Team, 2020) using the lme4 package (Bates et al., 2015). The model included Group, Rating Type, and a Group × Rating Type interaction as fixed effects, and by-speaker and by-listener random intercepts. We also included familiarity with L2 Spanish speech as a covariate. We used the emmeans package (Lenth, 2020) for post-hoc comparisons to locate statistically significant between-group differences. This package uses the Tukey method to account for multiple comparisons. As shown in Table 5, none of the between-group comparisons reached significance, suggesting that the three listener groups rated the speakers similarly on all three dimensions.

SIMULATING RELIABILITY AT DIFFERENT ONLINE RATER SAMPLE SIZES

We were also interested in how the reliability of data collected through AMT would change depending on the number of raters recruited. We reasoned that this information could help researchers make their studies more efficient by recruiting the number of

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

TABLE 5. Summary of least square means: group × rating type

	Comprehensibility		Fluency		Foreign accent	
	Estimate (*SE*)	*p*	*Estimate* (*SE*)	*p*	*Estimate* (*SE*)	*p*
Lab vs. L-Guided	0.19 (0.23)	.68	0.13 (0.23)	.82	0.04 (0.23)	.98
Lab vs. Any Dialect	0.17 (0.23)	.75	0.29 (0.23)	.43	0.16 (0.23)	.77
L-Guided vs. Any	0.02 (0.19)	.99	0.15 (0.19)	.72	0.20 (0.19)	.56

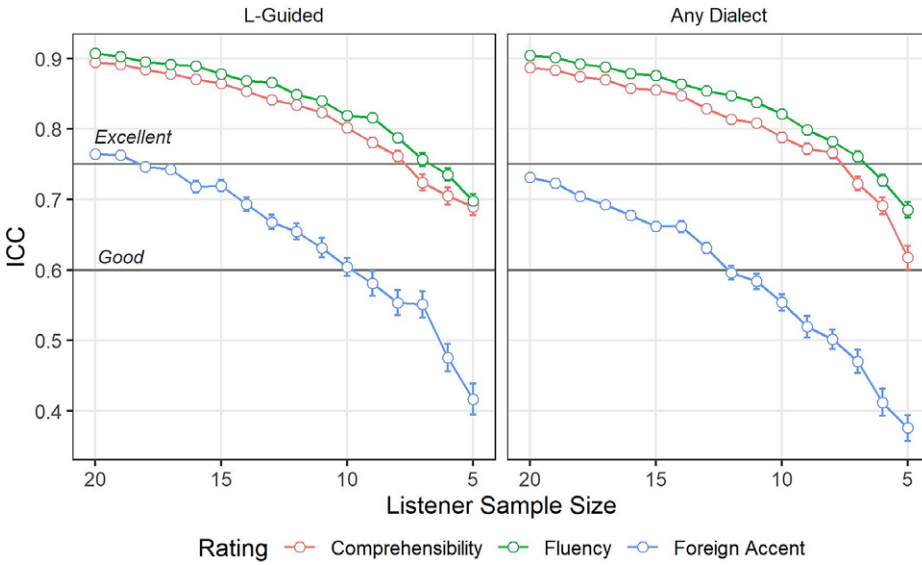

FIGURE 3. Reliability of the AMT ratings at different listener sample sizes.
Note. Each sample size simulation consists of 100 runs. These ICC estimates hold for the blocked design of the current study, in which listeners evaluated at least 11 L2 files and up to 22 files if they participated in both experimental blocks. According to Cicchetti (1994), ICC > .60 = good and ICC > .75 = excellent. Solid black lines have been added to the figure at these values.

listeners needed for reliability (while also considering issues of statistical power, depending on how the ratings are used). We generated 100 samples of *k* raters (e.g., *n* = 20, 19, 18 … 5) by randomly sampling raters within each group. For instance, at *n* = 20, we randomly sampled 20 raters from the AMT any-dialect group and 20 raters from the AMT learner-guided group, repeating this process 100 times to create 100 distinct rater groups. We estimated ICC(C, *k*) to examine the mean and range of ICCs observed at each rater sample size for each construct and AMT group. As shown in Figure 3, this simulation suggests that comprehensibility and fluency could be estimated with excellent reliability at sample sizes of seven to eight listeners and with good reliability with samples as small as five listeners given the other design features of this study (i.e., a blocked design in which each listener evaluates at least 11 items). On the other hand, a larger listener sample size would be required to obtain good reliability for the foreign accent ratings.

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

DISCUSSION

In this study, we set out to develop a more valid and robust approach to collecting L2 speech ratings online in AMT. We improved upon Nagle's (2019) method by including a screening task that we used to validate AMT workers. We also built timers into the interface that ensured that workers (1) listened to each audio file in its entirety before making their ratings and (2) moved through the task at a reasonable pace without backtracking. We scrutinized the reliability of the resulting data by (1) computing two-way consistency intraclass correlation coefficients, (2) fitting Rasch models to examine differences in rater severity and fit, and (3) fitting a mixed-effects model to examine between-group differences in scoring. We also simulated the reliability of the ratings data at different listener sample sizes to provide preliminary insight into the number of online listeners required to produce good to excellent scale reliability. In all of these analyses, we compared two AMT listener sampling strategies: sampling listeners from the dialects to which FL Spanish learners had been exposed through their instructors and sampling listeners from a broad range of dialects without considering the input FL learners had received. We compared these two groups to a group of US-based laboratory raters who were recruited locally at the university where the research took place.

Overall, the results showed that the data collected from all three groups, when aggregated, was highly reliable. Reliability estimates were in the excellent range for comprehensibility and fluency and in the acceptable to good range for foreign accent. Thus, the results of this study corroborate Nagle's (2019) findings and provide further evidence that comprehensibility and fluency ratings can be collected reliably online. The fact that reliability was lower for the foreign accent scale is not entirely surprising. Many studies that use L2 speech ratings as a global pronunciation outcome measure focus on degree of accentedness in reference to a local variety of the L2. In that case, ratings may exhibit higher reliability because listeners have the same internal anchor point. In contrast, in the present study, listeners were asked to evaluate foreign accent, judging the sample not in relation to a single local variety but in relation to any native variety of the L2. Such an approach entails that listeners understand what speech characteristics would surface in nonnative speech versus those that might surface in another native variety of the L2.

On the post-task survey, AMT workers indicated that they did not have difficulty completing the ratings (M = 83.88/100) and that they had understood the foreign accent and comprehensibility scales well (M = 93.08 and 93.09, respectively). At the same time, their open-ended comments suggested that they were sensitive to scaling issues, particularly with respect to foreign accent. For instance, one worker commented that "it would be helpful to give an example of each scale step because each listener will give different ratings depending on their perspective." Another said, "The foreign accent part needs some explanation. Since it's almost a binary answer, the intent isn't clear." This comment in particular signals that our attempts to orient listeners toward nonnative accents on the foreign accent scale (e.g., by including wording that assigning the best score indicated that the speaker could be a native speaker of Spanish) was not entirely successful. Providing additional files for evaluation at the screening stage along with feedback and/or more robust scale descriptors could help mitigate these concerns.

Reliability analyses revealed surprisingly few between-group differences. Reliability coefficients were similar across the board, especially for comprehensibility and fluency,

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

and the mixed-effects model and post-hoc comparisons revealed no significant differences in mean scores. There were, however, two areas where the learner-guided sampling strategy seemed to outperform the any-dialect strategy. First, Rasch modeling showed fewer statistically distinct levels of rater severity in the learner-guided group than in the any-dialect group (4 vs. 6), which could be interpreted as a sign of greater consistency among raters who were sampled from dialect regions that represented significant sources of input for the FL speakers included in this study. Second, reliability simulations at a variety of sample sizes suggested that higher reliability could be obtained for learner-guided samples, a trend that became more pronounced in the smallest sizes that we simulated. Yet, this finding deserves qualification for two reasons. First, the learner-guided sampling strategy could have yielded higher reliability simply because fewer dialects were sampled, making that group inherently less variable than the any-dialect group. Second, and to that point, although workers from Argentina, Mexico, and Spain were recruited at the screening stage and received the study-specific qualification granting them access to the experimental task, ultimately, most of the workers who completed the experimental task were from Spain, which could also account for the higher reliability observed for the learner-guided listener group. Put another way, the learner-guided group represented a narrow range of Peninsular Spanish dialects. This, coupled with the fact that approximately half of the speakers reported that they had taken Spanish courses from instructors who were native speakers of Peninsular Spanish, likely accounts for the differences between the two sampling strategies. Overall, then, the present findings do not necessarily show that learner-guided sampling is a superior sampling strategy, but they do inspire confidence in a variety of approaches to online listener recruitment.

Of course, in addition to reliability, listener sampling practices should be informed by conceptual considerations. For example, if the goal is to help FL learners communicate successfully with a specific group of individuals with whom they will interact in the future (e.g., when they study or intern abroad), then, to the extent possible, listeners should be recruited from that group. Future research should test that approach, which might prove especially useful for upper-level language students who have cultivated a deeper understanding of why and with whom they plan on using the L2. Admittedly, targeting a very specific group of listeners may be difficult through online platforms. For one, the only geographic filter available in AMT at the time of testing was a country-level filter. Recruiting individuals from the same country to serve as listeners would likely result in a narrower range of target varieties, as previously discussed, but it would not guarantee that all listeners speak the same variety of the target language because there is often substantial dialectal variation within a single country or region.

In addition to our primary goal of developing a more robust interface and checking the reliability of the resulting data, our secondary objective was to understand the demographic characteristics of AMT workers who were nonnative English speakers (in this case, Spanish speakers). Descriptively, the AMT workers we recruited were bi- or multilingual individuals with moderate to high proficiency in English who were accustomed to interacting with L2 Spanish speakers in both personal and professional contexts. For the most part, they were also university-educated; most had completed a 4-year degree, and many had an advanced degree in their field. It is also clear that listeners were technologically literate. This group, therefore, represents one important subset of potential interlocutors that researchers can access online.

RECOMMENDATIONS FOR DOING ONLINE SPEECH RESEARCH

The findings from this study have implications for doing online speech research. First, they underscore the necessity of including a screening task, which allows researchers to validate participants' work and ensure that they meet inclusion criteria for the study. Screening data can also provide researchers with insight into parts of the task or interface that are not functioning properly or that need further clarification. Another advantage of screening tasks is that they allow the researcher to begin creating a database of vetted workers who can be authorized to complete similar tasks in the future. The present study also confirms the utility of implementing a posttask survey to diagnose problems with the user interface and instructions. For instance, some workers indicated that they had trouble interpreting the foreign accent scale, which likely contributed to its lower overall reliability compared with the other two rated dimensions. Based on such feedback, the task could be updated in a future iteration to make that scale clearer, such as by providing additional descriptive information, examples, and so on.

As with any study, some aspects of methodology must be specified clearly from the start, such as participant inclusion and exclusion criteria. In this study, we were fairly lenient in terms of inclusion criteria: participants had to indicate that they were a native Spanish speaker and that they were born in a Spanish-speaking country, and they had to use the rating scales properly when evaluating the sample audio files that were part of the prescreening task. These criteria were easy to implement through geographic and study-specific filters. Geographic filters can be useful for recruiting workers from a certain region, but those filters guarantee only that workers presently reside in that region. Thus, it is important not to make assumptions about other demographic characteristics on the basis of residence (or, more precisely, the location of the user's IP address). For example, in the present study, one of the listeners included in the AMT L-Guided group indicated that he was born in Venezuela but was living in Spain at the time. Although that listener would undoubtedly be familiar with the characteristics of Peninsular Spanish, it would be inaccurate to classify him as a native speaker of that variety. Researchers should carefully consider how they classify workers based on demographic variables such as country of residence, country of origin, and so on, as well as how they use those variables to compose listener groups. Worth noting is that AMT offers researchers a variety of flexible options for implementing other inclusion and exclusion criteria, using both prebuilt AMT filters (some of which are associated with an additional fee) and in-house/study-specific filters that researchers can create.

On a related note, in this study, we created a learner-guided group by recruiting raters from Argentina, Mexico, and Spain. After screening an initial group of raters, we made the experimental task available to the entire group. The unintended consequence of this decision was that most of the raters who completed the experimental task were from Spain. Another option would have been to deploy the task separately to each of those regions, in which case we would have been able to end up with, for instance, 10 listeners from each country (for an example, see Huensch & Nagle, 2021). Thus, we recommend that researchers consider whether it would be necessary and/or advantageous to deploy the experimental task multiple times to target several different listener groups, leading to a more balanced and representative listener group. The fact that most AMT L-Guided listeners were from Spain also

https://doi.org/10.1017/S0272263121000292 Published online by Cambridge University Press

underscores the dynamics of AMT, and, indeed, the dynamism of the AMT userbase. In some countries, such as Spain, the userbase seems to be quite large, and users seem to log on and complete tasks quite frequently, whereas in others, it may be difficult to recruit a sufficient number of workers. What's more, the userbase is constantly evolving, as new workers join the platform and existing workers leave it. It is, therefore, unclear if AMT can be used for more complex, repeated-measures research designs. Future work should address this topic and should also explore the utility of AMT for collecting other types of L2 data.

Last but certainly not least, researchers must consider the ethical dimension of online research, which includes considering who has the means to participate. Clearly, workers must have access to a device and a reliable Internet connection, which necessarily excludes a large number of individuals who lack access to one or both. It is also important to acknowledge that, in many countries, institutional review boards have developed policies and requirements for doing online research, including policies that specifically address AMT. For instance, in some cases, researchers may be asked to make participants aware of the fact that their data may be stored in jurisdictions where governments have expanded access to personal records such as IP addresses. Finally, one weakness of AMT is that it does not offer researchers and workers a convenient means of dialoguing with one another during the research process. This means that researchers must be intentional about including task elements, such as feedback forms, that allow workers to offer suggestions, voice concerns, and, if necessary, lodge complaints and notify researchers of adverse effects.

CONCLUSION

Online and distance research methods are becoming increasingly common. Online research comes with challenges, but it also offers some advantages over an in-person approach. For one, it can broaden potential participant pools. It also allows researchers to carry out studies that otherwise would be difficult or impossible to execute. Such is the case for L2 speech ratings. When there are no local listeners to recruit, either because there are few local listeners of the L2 or because local listeners do not match the target group that researchers need for their study, online recruitment and data collection can be a viable and even desirable alternative. The results of this study show that online ratings can be as reliable as those collected in person. This study also raises important questions about how raters should be recruited and how constructs should be adapted and defined unambiguously in a new research context. Ultimately, these questions can only be answered in light of other methodological considerations. What is certain, however, is that online data collection is here to stay and will likely become more prevalent in an increasingly digital world. Future work should, therefore, replicate the current procedure using a larger number of samples provided by speakers of varying proficiency before broad conclusions can be reached regarding online speech rating procedures. It would also be fruitful to explore the extent to which other types of data (e.g., writing, speech) can be reliably and ethically collected online. In short, there is far more work to be done in this area, including work targeting crowdsourcing platforms other than AMT.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0272263121000292.

COMPETING INTERESTS

The authors declare none.

NOTES

[1]Compared with picture description, where speakers receive a set of images and are tasked with generating a story to fit what they see, we reasoned that retelling a narrative without having to imagine the plot would level the playing field in terms of the amount and complexity of content that speakers produced, making the task easier for the intermediate L2 speakers we included in the study.

[2]Ten speakers indicated that most of their Spanish teachers had been native speakers. The remaining 13 speakers indicated that throughout primary and secondary school most of their instructors had been native English speakers whereas in college many, if not all, had been native Spanish speakers. When asked to recall to the best of their ability the dialects to which they had been exposed, speakers listed the following varieties of Spanish: Argentinian ($n = 20$), Peninsular ($n = 10$), Mexican ($n = 9$), Puerto Rican ($n = 1$), and Peruvian ($n = 1$) Spanish.

[3]We chose not to implement any catch trials (e.g., trials in which a low amplitude sound is presented that would be difficult to detect without the use of headphones in a quiet environment) to confirm that online listeners were wearing headphones to make the task more efficient. Thus, although we asked online listeners to use headphones, we cannot verify that they did so. Even if they opted not to use headphones, it seems unlikely that failure to do so would have a significant impact on results because speech ratings have been reliably collected in a group setting where some ambient noise is likely to be present.

REFERENCES

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01.

Bergeron, A., & Trofimovich, P. (2017). Linguistic dimensions of accentedness and comprehensibility: Exploring task and listener effects in second language French. *Foreign Language Annals*, *50*, 547–566. https://doi.org/10.1111/flan.12285.

Callison-Burch, C., & Dredze, M. (2010). Creating speech and language data with Amazon's mechanical Turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical Turk* (pp. 1–12). Los Angeles, California.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290. https://doi.org/10.1037/1040-3590.6.4.284.

Derwing, T. M., & Munro, M. J. (2009). Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *Canadian Modern Language Review*, *66*, 181–202. https://doi.org/10.3138/cmlr.66.2.181.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, *63*, 163–185. https://doi.org/10.1111/lang.12000.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement*. Peter Lang.

Foote, J. A., & McDonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation*, *3*, 34–56. https://doi.org/10.1075/jslp.3.1.02foo.

Fort, K., Adda, G., & Bretonnel Cohen, K. (2011). Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, *37*, 413–420. https://doi.org/10.1162/COLI_a_00057.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*, 213–224. https://doi.org/10.1002/bdm.1753.

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*. Advance online publication. 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*, 135–159. https://doi.org/10.1080/15434303.2013.769545.

Isaacs, T., Trofimovich, P., & Foote, J. A. (2017). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, *35*, 193–216. https://doi.org/10.1177/0265532217703433.

Kennedy, S., Foote, J. A., & Dos Santos Buss, L. K. (2015). Second language speakers at university: Longitudinal development and rater behaviour. *TESOL Quarterly*, *49*, 199–209. https://doi.org/10.1002/tesq.212.

Lenth, R. V. (2020). Emmeans: Estimated marginal means, aka least-squares means. R package version 1.5.2-1. https://CRAN.R-project.org/package=emmeans.

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

Loukina, A., Lopez, M., Evanini, K., Sundermann-Oeft, D., Ivanov, A. V., & Zechner, K. (2015). Pronunciation accuracy and intelligibility of non-native speech. In *Interspeech 2015* (pp. 1917–1921). Dresden, Germany.

Martin, D., Hanrahan, B. V., O'Neill, J., & Gupta, N. (2014). Being a turker. In *CSCW 2014 - Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 224–235). Association for Computing Machinery. https://doi.org/10.1145/2531602.253166.

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). Taylor & Francis.

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.

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*, 1–19. 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*, *3*, 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*, 329–351. https://doi.org/10.1075/jslp.20009.nag.

O'Brien, M. G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning*, *64*, 715–748. https://doi.org/10.1111/lang.12082.

O'Brien, M. G. (2016). Methodological choices in rating speech samples. *Studies in Second Language Acquisition*, *38*, 587–605. https://doi.org/10.1017/s0272263115000418.

Paolacci, G., & Chandler, J. (2014). Inside the Turk. *Current Directions in Psychological Science*, *23*, 184–188. https://doi.org/10.1177/0963721414531598.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Peabody, M. A. (2011). *Methods for pronunciation assessment in computer aided language learning*. Massachusetts Institute of Technology.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon mechanical Turk. *Behavioral Research Methods*, *46*, 1023–1031. https://doi.org/10.3758/s13428-013-0434-y.

R Core Team (2020). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in mechanical turk. *Paper presented at the CHI '10 extended abstracts on human factors in computing systems*. Atlanta, GA.

Saito, K., Dewaele, J.-M., Abe, M., & In'nami, Y. (2018). Motivation, emotion, learning experience, and second language comprehensibility development in classroom settings: A cross-sectional and longitudinal study. *Language Learning*, *68*, 709–743. https://doi.org/10.1111/lang.12297.

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

Saito, K., Trofimovich, P., & Isaacs, T. (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.

Southwood, M. H., & Flege, J. E. (1999). Scaling foreign accent: Direct magnitude estimation versus interval scaling. *Clinical Linguistics & Phonetics*, *13*, 335–349. https://doi.org/10.1080/026992099299013.

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

Trofimovich, P., Lightbown, P. M., Halter, R. H., & Song, H. (2009). Comprehension-based practice. *Studies in Second Language Acquisition*, *31*, 609–639. https://doi.org/10.1017/s0272263109990040.

Trofimovich, P., Nagle, C. L., O'Brien, M. G., Kennedy, S., Taylor Reid, K., & Strachan, L. (2020). Second language comprehensibility as a dynamic construct. *Journal of Second Language Pronunciation*, *6*, 430–457. https://doi.org/10.1075/jslp.20003.tro.

Wang, H., Qian, X., & Meng, H. (2013). Predicting gradation of L2 English mispronunciations using crowdsourced ratings and phonological rules. In P. Badin, T. Hueber, G. Bailly, D. Demolin, & F. Raby (Eds.), *Proceedings of speech and language technology in education (*SLaTE 2013*)* (pp. 127–131). Grenoble, France.

Wisniewska, N., & Mora, J. C. (2020). Can captioned video benefit second language pronunciation? *Studies in Second Language Acquisition*, *42*, 599–624. https://doi.org/10.1017/s0272263120000029.