Developing and validating a methodology for crowdsourcing L2 speech ratings in Amazon Mechanical Turk

Charles Nagle

Iowa State University

Abstract

Researchers have increasingly turned to Amazon Mechanical Turk (AMT) to crowdsource speech data, predominantly in English. Although AMT and similar platforms are well positioned to enhance the state of the art in L2 research, it is unclear if crowdsourced L2 speech ratings are reliable, particularly in languages other than English. The present study describes the development and deployment of an AMT task to crowdsource comprehensibility, fluency, and accentedness ratings for L2 Spanish speech samples. Fifty-four AMT workers who were native Spanish speakers from 11 countries participated in the ratings. Intraclass correlation coefficients were used to estimate group-level interrater reliability, and Rasch analyses were undertaken to examine individual differences in rater severity and fit. Excellent reliability was observed for the comprehensibility and fluency ratings, but indices were slightly lower for accentedness, leading to recommendations to improve the task for future data collection.

Keywords: research methods; speech ratings; Spanish; reliability; many-facet Rasch measurement

1. Introduction

Second language (L2) pronunciation research routinely involves recruiting listeners to rate various aspects of L2 speech. These ratings serve as the basis for investigating issues ranging from how learners' speech develops over time (Derwing & Munro, 2013; Kennedy, Foote, & Dos Santos Buss, 2015) to the linguistic factors that undergird listeners' perception of speakers' comprehensibility, fluency, and accentedness (O'Brien, 2014; Saito, Trofimovich, & Isaacs, 2017; Trofimovich & Isaacs, 2012). Yet locating, recruiting, and scheduling listeners can prove challenging, particularly for researchers working on less commonly taught languages or in contexts where native speakers are scarce.

By connecting researchers with a larger and more diverse pool of listeners, crowdsourcing platforms such as Amazon Mechanical Turk (AMT) offer a potential solution to some of the practical barriers associated with rater recruitment. As crowdsourcing has become an increasingly recognized means of data collection (Eskénazi, Levow, Meng, Parent, & Suendermann, 2013), researchers have recruited AMT workers to evaluate mispronunciations in native (McAllister Byun, Halpin, & Szeredi, 2015) and L2 speech (Peabody, 2011; Wang, Qian, & Meng, 2013), and to rate samples for features such as accentedness (Kunath & Weinberger, 2010). In addition to providing access to a larger and possibly more representative sample of listeners across a range of languages, services like AMT may prove to be a necessary data collection tool for designs that generate a large number of samples, such as the longitudinal pronunciation studies that are becoming more common in L2 speech research (e.g., Derwing & Munro, 2013; Nagle, 2018b; Saito, Dewaele, Abe, & In'nami, 2018).

For these reasons, two aspects of crowdsourced L2 speech ratings deserve more precise methodological attention. First, it is unclear whether crowdsourced L2 data is as reliable as data collected in a laboratory setting, and second, questions remain about the viability of collecting data in other languages since most studies have focused on English (for a notable exception, see Gelas, Teferra Abate, Besacier, & Pellegrino, 2011). Addressing these gaps, the present study reports on (1) the development and deployment of an AMT rating template used to crowdsource speech ratings for L2 Spanish; (2) data collection, processing, and trimming based on responses to two control measures included to ensure that workers remained attentive throughout the task; (3) data reliability, assessed using intraclass correlation coefficients; and (4) rater and scale performance, evaluated through Rasch analysis.

2. Background

2.1 Online and laboratory L2 speech ratings

AMT is a crowdsourcing platform that allows requesters to divide a large project into a set of human intelligence tasks (HITs) that workers, or “Turkers,” can complete in exchange for a small amount of compensation. HITs are typically discrete, repetitive tasks that cannot be automated using artificial intelligence, such as tagging images with keywords or transcribing audio files. AMT empowers requesters with a variety of control measures that enable them to target certain subsets of the AMT user base, such as high-reputation workers—workers who have an overall HIT approval rating above 95%—or workers whose IP addresses are located within a certain geographic region. Requesters can also create specialized selection criteria that fit their needs, such as a language proficiency test (AMT also offers language certifications, but it is unclear how the tests were developed) or a training block to familiarize workers with the task.

In laboratory studies, researchers meet with listeners to collect demographic information, review instructions, and oversee a brief training or familiarization block. In contrast, AMT workers participate remotely, which means that researchers have limited insight into what workers are actually doing while completing the task, including how well they have understood instructions and how attentive they remain throughout the session. Consequently, two types of studies have been undertaken to examine the reliability of AMT data. First, researchers have evaluated AMT users' responses to a variety of individual difference measures, finding that AMT workers and local, laboratory participants display similar cognitive profiles, at least to the university undergraduates that most often participate in lab studies (Buhrmester, Kwang, & Gosling, 2011; Goodman, Cryder, & Cheema, 2013; Paolacci & Chandler, 2014; Paolacci, Chandler, & Ipeirotis, 2010). Directly relevant to the present research is a second set of studies reporting on the reliability of crowdsourced L2 speech data. Table 1 summarizes the methodological features and reliability indices of laboratory and AMT speech research for a representative sample of studies. As is apparent, in laboratory studies employing fully-crossed designs (i.e., all raters evaluate all speakers), reliability is most often assessed using Cronbach's alpha or the intraclass correlation coefficient, and good to excellent reliability (measure > .80) is observed in nearly all cases. Conversely, computing reliability for crowdsourced data sets is more complicated since unbalanced designs (i.e., a random set of *n* raters evaluates each speaker) are common. For example, Peabody (2011) developed an extension of Cohen's kappa to evaluate agreement among over 10,000 rater pairs, excluding pairs that did not rate at least ten of the same sentences. The aggregated kappa revealed moderate (κ = .51) agreement.

To mitigate the potential reliability issues inherent to data collection in AMT, scholars have devised quality control strategies such as attention checks to detect inattentive responders (Goodman et al., 2013; Paolacci et al., 2010). A common iteration of an attention check involves embedding explicit instructions on how to respond to a trial within the trial itself, such that only attentive workers will respond correctly. However, such posthoc screening practices have been criticized since they not only assume a constant level of attention across the task, but may also alter sample characteristics (Paolacci & Chandler, 2014). Moreover, other prescreening procedures, such as limiting HITs to high-reputation workers, have been shown to be equally effective. For instance, Peer, Vosgerau, and Acquisti (2014) found that attention check questions improved the reliability of data provided by low-reputation workers, but data from the high-reputation group displayed high reliability across a range of measures (e.g., Cronbach's alpha, central tendency bias) irrespective of the attentional manipulation. Recruiting high-reputation workers to complete tasks in or involving English is relatively straightforward given the large number of English-speaking workers based in the United States and India (Paolacci et al., 2010; Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010). In contrast, recruiting such workers in other languages with a smaller user base may be more difficult.

Table 1. Methodological features and reliability of laboratory and AMT L2 speech studies.

Study	L1	L2	Raters / AMT Workers	Sample	Scale	Reliability
Laboratory Studies						
Akiyama & Saito (2017)	English Eng./Japanese Eng./Chinese Chinese	Japanese	4 NS who were graduate students in linguistics	Two 30 second clips (pre/post design)	1000 point	Cronbach's α Comp. = .82
Bergeron & Trofimovich (2017)	Spanish	French	20 NS who were pursuing degree in education or pedagogy; half with knowledge of Spanish	30 second clip	1000 point	Cronbach's α Comp. = .91–.94 Accent. = .94–.95
Crowther et al. (2015)	Chinese Hindi-Urdu Farsi	English	10 NS who had completed or were enrolled in applied linguistics grad. program	30 second clip	1000 point	Cronbach's α Comp. = .86 Accent. = .93
Derwing & Munro (2013)	Mandarin Slavic	English	34 Monolingual NS and 10 highly proficient NNS	20-25 second clip at each time point (3)	9 point	Intraclass correlation (ICC) NS (34) / NNS (10) Comp. = .96 / .87 Fluency = .97 / .93 Accent. = .95 / .90
Isaacs & Thomson (2013)	Mandarin Slavic	English	40 NS who were experienced ESL teachers (20) or graduate students in other fields (20)	20 second clip	5 point 9 point	Cronbach's α Comp. = .92–.95 Fluency = .92–.94 Accent. = .94–.95 Kendall's W Comp. = .39–.50 Fluency = .41–.49 Accent. = .47–.53
Munro & Derwing (1995)	Mandarin	English	18 NS with knowledge of articulatory phonetics	3 4–17 word excerpts	9 point	ICC Comp. = .96 Accent. = .98

Nagle (2018)	English	Spanish	18 NS of various dialects of Spanish who were advanced speakers of L2 English	5 sentences at each time point (5)	9 point	ICC: two-way, consistency, average-measure Comp. = .93 Accent. = .94
O'Brien (2014)	English	German	25 L1 English speakers who were learners of L2 German of varying proficiency	20 second clip	9 point	ICC (nns samples only) Comp. = .22 Fluency = .08 Accent. = .15
AMT Studies						
Kunath & Weinberger (2010)	Arabic Mandarin Russian	English	50 workers located in the US (possibly including NNS)	Clips from the Speech Accent Archive	5 point	No report
Peabody (2011)	Cantonese	English	463 high-reputation workers located in the US	Sentences from CU-CHLOE corpus	3 point	Aggregated Cohen's κ to assess pairs of workers who evaluated the same sentences for mispronounced words Mispronunciation = .51
Wang, Qian, & Meng (2013)	Cantonese	English	287 workers	Sentences from CU-CHLOE corpus	4 point	Worker rank algorithim (based on a page rank algorithim for web pages) that incorporates Cohen's κ 190 workers retained

Note. NS = native speaker; NNS = non-native speaker; Comp. = comprehensibility; Accent. = accentedness; CU-CHLOE = Chinese University Chinese Learner of English. The laboratory studies employed a balanced, or fully-crossed, design in which all raters evaluated all speakers. The AMT studies employed an unbalanced, or random raters, design in which a group of *n* raters evaluated each speaker, in most cases 3-5 raters per file.

2.3 Demographics of AMT workers

The demographic characteristics of AMT workers have received significant attention in the literature, not just because this information is needed to report sample characteristics, but also because workers' experiences must be taken into consideration to design HITs that are user-friendly and ethical, especially in terms of compensation. When interpreting demographic data, it is important to bear in mind that demographic studies administered via AMT reflect the user base at the time of data collection. Thus, while demographic data may not be representative of the current population of workers, it does shed light on broad trends in worker characteristics over time. For example, the results of Ross, Irani, Silberman, Zaldivar, and Tomlinson (2010) suggest that AMT workers are highly educated and may depend on AMT as a source of income (see also, Fort, Adda, & Bretonnel Cohen, 2011). In a more recent study, Martin, Hanrahan, O'Neill, and Gupta (2014) analyzed posts to Turker Nation, a website where workers can share their experiences with AMT. Analyses confirmed that many workers rely on the income they generate through AMT, and that US workers in particular were concerned with earning a fair wage comparable to the federal minimum of $7.25 per hour. The authors also found that workers were committed to promoting successful worker-requester interactions; they helped one another locate the best HITs, posted critical evaluations of requesters, and even helped requesters improve HITs when the opportunity arose.

To examine the language demographics of AMT, Pavlick, Post, Irvine, Kachaev, and Callison-Burch (2014) asked bilingual workers to report their native language and country of residence. Workers' self-reported language ability was subsequently validated by geolocating their IP address and by asking them to translate words from the target language into English. Translations were checked against a gold standard computed through Wikipedia articles, and individuals whose translations displayed the highest degree of overlap with Google translate were removed. The authors then compared the quality of translations inside and outside of regions where the language was likely spoken, and assessed speed of completion for the translation HITs. Over 3,000 workers completed the language survey, resulting in 35 languages with at least 20 speakers. English and the languages of the Indian subcontinent were the most commonly reported, but languages such as Spanish, Chinese, and Portuguese were also well represented. Among the latter three, Spanish and Portuguese were ranked as high quality languages based on the number of active in-region workers and their speed.

2.4 The current study

Accumulated findings for AMT research tentatively indicate that the data is reliable, though reliability may be slightly lower than comparable data collected in a laboratory context. At the same time, attention check measures can help ensure that AMT workers remain on task, potentially enhancing the reliability of crowdsourced L2 speech data. Likewise, studies suggest that workers find tasks that come with clear instructions and formatting more desirable and complete them more successfully. Nevertheless more detailed reliability studies are needed, particularly studies involving languages other than English where AMT could prove particularly fruitful for connecting researchers with participants in less commonly represented L2s. The overall objective of the present study was therefore to assess the feasibility of collecting L2 Spanish speech ratings through AMT and to evaluate the reliability of the data, including various aspects of rater performance. The following research questions guided the study:

1. What percentage of data collected via AMT is valid after preprocessing for the attention check and near-native control measures?

2. Are L2 speech ratings collected via AMT reliable, and how does the reliability of AMT data compare to laboratory data?
3. What individual differences in rater and scale performance are evident when Rasch models are fit to the AMT ratings data?
4. What relationships are evident between rater background variables and rater performance and between the background variables and the ratings data?

3. Method

3.1 Speech samples

The speech samples included in this study were part of an unpublished longitudinal data set examining L2 Spanish learners' pronunciation development over time. Speakers were 16 L1 English university students (13 females) who were enrolled in a third- (*n* = 10) or fifth-semester (*n* = 6) communicative Spanish language course at the time of recruitment. The mean age of onset was 12 years (*SD* = 3.25, range 7–18), and speakers had between five and six years of previous Spanish coursework on average (*M* = 5.49, *SD* = 2.64, range = 1–12).

Speakers completed a picture narration and a prompted response task. On the former, they received a six-frame story depicting a dog sneaking into a picnic basket and eating the meal that two children had prepared with their mother (cf. Muñoz, 2006). Six key words (e.g., *canasta,* 'basket') were provided to facilitate the narration. For the prompted response, speakers were asked to describe their daily routine in Spanish in as much detail as possible[1].

Speakers were recorded individually in a sound-attenuated booth using a Shure SM10A head-mounted microphone connected to a laptop computer through an XRL-to-USB signal adapter. They had one minute to prepare before recording each task but were not allowed to script a response. Speech samples were collected on three occasions over the academic year: just before the midterm of fall semester, at the end of fall semester, and at the end of spring semester. Due to participant attrition, 39 samples (session 1, *n* = 16; session 2, *n* = 12; session 3, *n* = 11) were available for each task.

Following previous research (e.g., Derwing & Munro, 2013), the first 30 seconds of each clip were sampled, excluding false starts and selecting an end-point coinciding with a natural break in the response. Excerpts were normalized to a peak intensity of 70 dB. In addition to the learner samples, four near-native speaker samples were included as anchor or control clips, and seven attention checks were included as a means of establishing that listeners remained attentive throughout the rating task. Attention checks were created by replacing the last 5-10 seconds of a clip with the voice of a male native speaker of Argentinian Spanish indicating the scores the clip should receive (e.g., Assign this clip the following ratings: comprehensibility, 1; fluency, 9; accentedness, 2). The native speaker voice was spliced into clips provided by L2 speakers who were not included in the target speech samples. In total, there were 50 clips to be rated per task.

3.2 Development and deployment of the AMT HITs

The template for the rating task was developed in AMT (the code for the template is available for download through the IRIS digital repository). The task displayed a set of collapsible instructions that (1) summarized the purpose of the experiment, (2) presented the three constructs to be rated, and (3) outlined other important task features. The operationalization of constructs followed Derwing and Munro (2013). Comprehensibility was defined as how easy or difficult the speech was to understand, and workers were made aware of the fact that they should assess the extent to which concentrated listening was required to understand the speaker. Fluency was broadly defined as the rhythm of the speech, that is, whether or not speakers expressed themselves with ease, without pausing, or paused frequently and seemed to experience difficulty. For this construct, workers were instructed to ignore grammar issues. Accentedness was operationalized as deviations from any native variety of Spanish, and workers were made aware of the distinction between comprehensibility and accentedness (i.e., a speaker may be very comprehensible, or easy to understand, and at the same time have a noticeable accent).

Ratings were conducted simultaneously using three 9-point Likert scales where 9 was the best score (e.g., for comprehensibility, 1 = *very difficult to understand* and 9 = *very easy to understand*). Given that workers were asked to make three judgments for each file, and due to the exigencies of the online context, the interface allowed the audio to be played up to three times before the embedded player disappeared. Thus, workers could listen to the sample once and evaluate all constructs, as in simultaneous ratings, or listen to the sample once per construct, as in a sequential rating paradigm (O'Brien, 2016). Instructions made it clear that workers should listen to the whole clip before rating it and that attention checks would be included, meaning workers would occasionally receive instructions on how to score a clip. Following the presentation of the instructions and scales, workers were asked to provide basic biographical data: their country of origin, age, gender, the highest level of education they had completed, native language(s), and additional languages known. Although this portion of the HIT remained active once completed, workers were informed that they only needed to provide biographical data once. Finally, an optional text entry box at the bottom of the HIT enabled workers to comment on the task.

This template was used to collect ratings for the picture narrative and prompted response separately. For the picture narrative, an image of the dog story was embedded into the task, and workers were told that they would evaluate a clip extracted from speakers’ responses. For the prompted response, workers saw the prompt that speakers were given and were likewise told that they would evaluate a brief clip. In both cases, workers were paid $0.10 per assignment (i.e., 10 cents per audio file). The HIT was set to expire in two weeks, and 20 unique workers were requested per file (i.e., each file was evaluated by 20 individuals). Workers' assignments were set to be approved automatically in one hour to compensate them in a timely manner.

In AMT, a .csv input file is required to link audio files to the HIT (i.e., to tell the interface where to search for the audio file). Separate HITs for the picture narrative and prompted response were deployed twice, each time with a different randomization of audio files, to collect ratings from a wide range of L1 Spanish listeners. Visibility of the HIT was limited to workers with an IP address in a Spanish-speaking country. In each case, 2,000 ratings were collected (20 workers × 2 tasks × 50 samples = 2,000 ratings). The first set of HITs was active for one day. Preliminary inspection of the data revealed that most participants were from Venezuela. Therefore, the HITs were redeployed, excluding Venezuelan IP addresses, to collect ratings from workers who were native speakers of other Spanish dialects. The second set of HITs was active for just over a week. In total, 4,000 ratings were collected across the two HIT deployments.

3.3 AMT Workers

Fifty-five unique AMT workers participated in the ratings. All workers completed a short biographical survey, described in detail below. Other than one worker who did not disclose her age, there was no missing data. One worker identified himself as a native speaker of Arabic born in Syria and was therefore removed from the data set. The other 54 raters were native Spanish speakers (15 females) whose age ranged from 20–52 (*M* = 32.83, *SD* = 8.18). Most workers had completed some amount of higher education, with a four-year college degree or equivalent being the most common (*n* = 35). Venezuela was the most frequent country of origin (*n* = 22), followed by Mexico (*n* = 10), Colombia (*n* = 8), and Spain (*n* = 5). Fifty-two participants reported some knowledge of English, and nine reported knowledge of a third or fourth language (French, *n* = 4; Italian, *n* = 3; German, *n* = 3; Portuguese, *n* = 2). For complete rater data, including the number of files evaluated and exclusion criteria, see the Appendix.

4. Results

4.1 Attention checks and near-native control samples

Nearly 90% of the AMT data was retained after processing the attention checks and near-native samples, which indicates that the vast majority of AMT workers had understood the rating instructions and had remained attentive throughout the rating task. The attention checks were included to detect individuals who did not listen to the entire clip or who may have been distracted while completing the task. Workers who responded incorrectly to more than two checks were excluded from the data set, but a single incorrect response was permitted since it could be attributed to selecting the wrong radio button by accident. Four workers responded incorrectly to more than two checks, and three of the four had failure rates above 90%, suggesting that they were not completely focused on the task or had not understood the instructions adequately. On average, these workers rated 67.75 audio files (*SD* = 40.48) or 271 files in total, which represents 6.78% of the total data set (271 / 4000). Twelve additional raters were excluded because they did not complete at least two attention checks, and so the quality of their responses could not be validated via this measure. In general, these were individuals who rated very few clips on average (*M* = 12.25, *SD* = 10.20, range = 1–34). A total of 147 ratings were discarded for this group, representing 3.68% of the total data set. Aggregating data from these two groups of excluded workers, 10.45% of responses were eliminated from the data set, which means that 89.55% of the data was validated and retained through the attention check measure.

Near-native speaker samples were included as a validity check on rater performance since raters should be capable of distinguishing intermediate speakers from a near-native control group. Near-native L2 speakers were chosen because they represent a more ecologically-valid comparison for the intermediate learners who provided the target clips in this study. Data from three workers (72 responses, or 1.80% of the data set) was excluded because they did not rate at least two near-native samples. Once these three workers were removed, averages for the learner and near-native speaker groups on both tasks suggest little overlap in scores. As reported in Table 2, means were always higher for the near-native speakers (above seven in all cases), who mostly received scores on the upper half of the 9-point scale (cf. spark plots). The differentiation between the two groups suggests that the workers had understood the instructions and did in fact make use of the entire scale. This is further supported by the fact that 10 of the original 54 native Spanish speakers left feedback on the task, describing it as fun, interesting, and dynamic.

Table 2. Descriptive statistics for the learner and near-native speakers by task.

	Picture Narration				Prompted Response			
	M	*SD*	*Range*	Spark	*M*	*SD*	*Range*	Spark
L2 Learners								
Comprehensibility	4.78	2.09	1–9		5.56	1.04	1–9	
Fluency	4.05	2.14	1–9		4.94	1.99	1–9	
Accentedness	2.87	1.96	1–9		3.19	2.01	1–9	
Near-native Speakers								
Comprehensibility	8.39	1.05	4–9		8.74	.61	6–9	
Fluency	8.24	1.17	4–9		8.64	.75	5–9	
Accentedness	7.47	2.22	1–9		7.53	1.95	1–9	

Note. Spark = a small line graph showing the distribution of scores on the 9-point rating scale.

4.2 Reliability

In addition to establishing that workers carried out the ratings as intended, it is also necessary to examine the extent to which they agree with one another. Reliability coefficients were consistently high, suggesting that workers evaluated speakers' comprehensibility, fluency, and accentedness similarly, or that there was a high degree of consensus among workers. For interval data, an intra-class correlation coefficient (ICC) is an appropriate measure of interrater agreement (Hallgren, 2012; McGraw & Wong, 1996; Shrout & Fleiss, 1979). The calculation of the ICC depends on three parameters. If the same group of raters evaluates all speakers (a fully-crossed design), then a two-way model is appropriate. This model allows for systematic variances attributable to specific raters or rater-by-speaker interactions to be taken into account. Conversely, if a random set of raters is sampled for each speaker, as might be the case in a larger study, then a one-way model is appropriate because systematic variances cannot be computed. Second, the researcher must decide if agreement or consistency is desirable. Consistency refers to whether the order of ratings observed for speakers holds across the data set (i.e., if all raters score speaker *a* higher than speaker *b*, etc.), and agreement refers to absolute agreement among raters. Lastly, the unit of generalization must be specified. If ratings are meant to generalize to the scores provided by a single individual rater, then a single-unit measure is appropriate. If ratings are meant to generalize to an average rating provided by a set of *n* raters, then an average measure is appropriate.

Two data sets were analyzed to examine interrater consistency. The first data set consisted of the total pool of 35 raters who were retained after the attention check and near-native sample screening procedures. Given that this data set was not fully crossed, six one-way, consistency, average-measures ICCs were computed to estimate reliability for each combination of construct (comprehensibility, fluency, and accentedness) and task (picture narration and prompted response). Fully crossed data sets for the picture narrative (*n* = 18) and prompted response ratings (*n* = 20) were subsequently constructed and analyzed for comparability with other L2 speech rating studies, most of which employ a fully-crossed design. For these data sets, interrater consistency was computed using six two-way, consistency, average-measures ICCs. As reported in Table 3, reliability was in the excellent range for comprehensibility and fluency and in the very good range for accentedness for both the unbalanced (i.e., one-way random model) and fully-crossed (i.e., two-way random model) data sets.

Table 3. Intraclass correlation coefficients and confidence intervals by construct and task.

	Comprehensibility		Fluency		Accentedness	
Model/Task	*ICC*	95% CI	*ICC*	95% CI	*ICC*	95% CI
Unbalanced (one-way)						
Picture	.92	[.88, .95]	.96	[.94, .98]	.89	[.83, .93]
Prompt	.90	[.85, .94]	.93	[.90, .96]	.87	[.80, .92]
Fully-Crossed (two-way)						
Picture	.92	[.88, .96]	.96	[.93, .97]	.88	[.82, .93]
Prompt	.90	[.85, .94]	.94	[.90, .96]	.89	[.84, .94]

Note. For one-way, consistency, average-measure ICC, *n* = 35. For two-way, consistency, average-measure ICC, *n* = 18 and 20, for picture narrative and prompted response, respectively. The average measure ICC reflects reliability based on aggregated data from *n* raters.

4.3 Rasch modeling

Traditional reliability metrics such as the ICC and Cronbach's alpha are group-level estimates that indicate whether a group of raters has evaluated speakers similarly, assigning speakers the same ratings (agreement or consensus) or ordering speakers similarly (consistency). As Eckes has observed, these "statistics often mask non-negligible differences within a group of raters" (2015, p. 66). For example, raters may vary in terms of leniency or may make use of a limited portion of the rating scale. It could also be the case that these forms of rater bias occur only when certain features are evaluated. Many-facet Rasch modeling is ideal for detecting and quantifying individual variation in rater performance, including the aforementioned rater effects (see Myford & Wolfe, 2003). This type of analysis can also be applied to other aspects of the rating procedure, such as scale performance (i.e., Was the length of the scale appropriate and were the steps sufficiently distinct and of the same approximate magnitude?).

One of the key features of Rasch is that all facets (e.g., raters, speakers, constructs, etc.) are simultaneously calibrated onto the same logit scale, which facilitates comparison among the different facets of the rating design. In an ideal scenario, speakers would exhibit a wide logit spread, indicative of a range of proficiencies on the relevant measure, and raters a narrow spread, indicative of similar levels of rater severity when evaluating the speakers. Rasch also computes unbiased estimates of speaker performance adjusting for differences in rater severity. Finally, unlike traditional measures such as Cronbach's alpha, which require a balanced or fully-crossed data set, Rasch can function with unbalanced data sets as long as the facets are sufficiently connected (i.e., as long as multiple raters have evaluated each speaker, in which case speakers and raters are sufficiently, but not necessarily fully, crossed). Consequently, Rasch modeling provides a more comprehensive account of within-group (intrarater) differences, allowing the researcher to pinpoint and improve upon problematic aspects of the rating procedure.

In this study, a many-facet Rasch analysis was undertaken to investigate the extent to which individual raters differed in severity (see, Eckes, 2005), and to gain insight into the structure and performance of the 9-point rating scales (Isaacs & Thomson, 2013). For each construct, a rating scale model was fit to the trimmed data set (n = 35)[3] with the following four facets: speaker, rater, time, and task. Among other findings, these analyses revealed that AMT workers exhibited variable levels of severity, especially when evaluating accentedness. Additionally, scale steps were compressed for all three constructs, which suggests that a 5- or 7-point scale may have been more appropriate given the relative homogeneity of speakers sampled.

4.3.1 Calibration of speakers, raters, time, and tasks. Summary model statistics are presented in Table 4. As can be seen, there were statistically significant differences in rater severity for all three constructs. The significant chi-square statistics indicate that at least two raters exhibited distinct levels of severity, the separation indices suggest between four (fluency) and five (comprehensibility and accentedness) severity strata, and the reliability coefficients demonstrate that differences in rater severity were reliable (for the raters facet, low reliability is desirable as it would indicate no significant differences in severity). Model statistics likewise indicate that speakers were reliably differentiated into approximately five to six levels of performance. Examination of logit spreads for speakers and raters revealed a wider spread for raters, which can be attributed to the relative size of each facet (*n* = 11 for speakers who completed all three sessions vs. *n* = 35 for raters) and the homogeneity of the intermediate learners of L2 Spanish sampled in the present study. Logit spreads of .92, 1.03, and 1.76 were observed for speakers' comprehensibility, fluency, and accentedness compared to 1.69, 1.35, and 2.03 spreads for raters.

For the task facet, reliable differences were observed across the board, though greater differentiation of the picture narration and prompted response tasks was evident for comprehensibility and fluency (.32 logit spread for both) than for accentedness (.14 logit spread). The picture narration was more difficult than prompted response, with speakers receiving higher scores on average on the prompted response. The chi-square statistic for the time facet reached significance for comprehensibility and fluency but not for accentedness. Two levels were reliably calibrated for comprehensibility, and observed and fair averages for each session suggest that comprehensibility scores were slightly higher on average at the third session Three levels were calibrated for fluency, suggesting that scores for fluency increased incrementally from the first to the third session.

Table 3. By-construct summary statistics for the many-facet Rasch models.

Statistics	Speakers	Raters	Time	Tasks
Comprehensibility				
M	.00	.00	.00	.00
M SE	.05	.08	.02	.02
χ^2	327.00*	844.50*	10.00*	143.60*
df	10	34	2	1
Separation index	5.65	4.95	2.01	11.94
Separation reliability	.97	.96	.80	.99
Fluency				
M	–.34	.00	.00	.00
M SE	.04	.08	.02	.02
χ^2	441.60*	590.40*	33.40*	163.60*
df	10	34	2	1
Separation index	6.50	3.92	3.79	12.75
Separation reliability	.97	.94	.94	.99
Accentedness				
M	–.85	.00	.00	.00
M SE	.05	.09	.02	.02
χ^2	319.90*	922.80*	.90	26.40*
df	10	34	2	1
Separation index	5.38	5.19	.00	5.04
Separation reliability	.97	.96	.00	.96

Note. * indicates $p < .05$.

4.3.2 Rater fit. Rater fit was evaluated by examining the infit mean-square statistic for each individual worker. A mean-square value of 1 indicates a perfect fit to model expectations, whereas values below 1 indicate overfit (less variation than expected) and values above 1 misfit (more variation than expected). Linacre (2002) recommended a lower limit of .50 for overfit and an upper limit of 1.50 for misfit, and scholars have suggested that the latter is more problematic than the former (Eckes, 2015; Myford & Wolfe, 2003). In the present analysis, overfit and misfit limits were narrowed slightly to .60 and 1.40 to account for the sample size of the rater facet (Wu & Adams, 2013). Table 5 reports the number and percentage (in parentheses) of raters falling into each category for each construct. As displayed, raters were fairly consistent in their use of the comprehensibility and fluency scales, since in both instances there were few cases of misfit.

In contrast, the fact that eight raters were misfitting for accentedness suggests that a substantial proportion of raters may have adopted an idiosyncratic rating strategy.

Table 5. Number and percentage of raters classified according to fit for each rating scale.

Fit range	Comprehensibility	Fluency	Accentedness
Overfit (< .60)	4 (11.42)	0 (0)	3 (8.57)
Fit (.60 – 1.40)	27 (77.14)	32 (91.43)	24 (68.57)
Misfit (> 1.40)	4 (11.42)	3 (8.57)	8 (22.86)

4.3.3 Rating scale use and structure. According to Eckes (2015), indicators of scale quality include a regular distribution of frequencies across categories, a monotonic increase in average measures across categories, outfit mean-square values below 2.0 for each category, and 1.40–5.00 logit steps between categories. For comprehensibility, response frequencies exceeded 10% for categories 3–8, with category 7 selected the most often (17%) and categories 1 and 9 selected far less frequently (4%). The scale increased monotonically across all categories, and outfit mean-square statistics were acceptable, ranging from .80 for category 7 to 1.20 for category 2. However, category thresholds did not increase by at least 1.40 logits per step, which is not surprising given the relative homogeneity of the speaker facet. The smallest threshold step was .06 logits from category 4 to 5 and the largest .68 from category 7 to 8. To make categories more distinct, especially when speakers' proficiency levels are comparable, a 5- or 7-point scale could be employed. Similar results were obtained for fluency. Raters selected categories 2 through 7 with approximately the same frequency (11-15%), and categories 1 and 8 also displayed nontrivial frequencies of 9% and 7%, respectively. In contrast, category 9 was employed in only 2% of cases and was the only category for which the scale measure did not increase monotonically. The low frequency of the category could account for the fact that it did not continue the trend of increasing values with each scale step. Despite the reversal between categories 8 and 9, the outfit mean-square statistic for the latter (1.10) was within the acceptable range. Distances between category thresholds fell below the recommended value of 1.40 logits, indicating that at least some of the categories could be combined to create a shorter scale.

In contrast to comprehensibility and fluency, examination of category frequencies and thresholds for the accentedness scale revealed quality issues. Frequencies were substantially skewed toward the lower categories, ranging from 26% of responses for category 1 to 11% of responses for category 4. The cumulative frequency for categories 1–4 was 76% compared to 24% for categories 5–9. As would be expected based on the frequency data, thresholds were considerably compressed, including a reversal of categories 5 and 6 (–.13 for the former vs. –.20 for the latter). Consequently, rater fit and scale use data suggest that the accentedness scale should be revised.

4.4 Rater characteristics: age, gender, and education

Workers were asked to report basic biographical data: age, gender, country of origin, and level of education (four levels: high school, bachelor’s degree or equivalent, master’s degree, or doctoral degree). Two analyses were undertaken related to worker characteristics. First, patterns of rater misfit and overfit were descriptively analyzed to determine if problems with rater fit could be attributed to any of the background variables. Second, age, gender, and level of education completed were included as fixed effects in mixed-effects models of comprehensibility, fluency, and accentedness using the trimmed (*n* = 35) data set. Level of education was recoded into a categorical variable contrasting individuals who had completed a bachelor’s degree (*n* = 23) with individuals who had completed a graduate degree (*n* = 9). It was not possible to include high school degree due to the small cell size for that category (*n* = 4), nor was it possible to evaluate L1 dialect given unequal sample sizes across cells. Random intercepts for raters were included in all three models.

As displayed in Table 6, issues with rater fit cut across the four demographic variables collected in this study. The fact that workers 5 and 42 were categorized as misfitting on all three constructs suggests that in each case they applied a different strategy than the rest of the workers. Thus, removing the data these workers provided could be warranted. When mixed-effects models were fit to the rater data, no significant relationships were evident among the background characteristics and the comprehensibility and fluency ratings. However, individuals with a graduate degree tended to assign learners higher scores for accentedness (estimate = .85, *SE* = .39, *p* = .04), which indicates that they perceived them as being slightly less accented.

Table 6. Worker background characteristics and misfit/overfit classification.

Worker	Age	Gender	Country	Education	Misfit	Overfit
5	35	F	Venezuela	Bachelor	C / F / A	
7	25	F	Venezuela	Bachelor	A	
8	31	M	Spain	Bachelor	A	
10	26	M	Venezuela	Bachelor	C / A	
12	52	M	Venezuela	Bachelor		C / A
20	21	M	Mexico	Bachelor		C
22		F	Venezuela	Bachelor		C / A
40	42	M	Mexico	High School	A	
41	27	M	Colombia	High School	C / F	
42	28	F	Colombia	PhD	C / F / A	
44	24	M	Colombia	Masters		
45	36	M	Honduras	Bachelor		C / A
46	48	M	El Salvador	Bachelor	A	
47	31	M	Mexico	Bachelor	A	

Note. Worker 22 did not report her age. C = comprehensibility; F = fluency; A = accentedness.

5. Discussion

In this study, a template was developed to collect L2 Spanish speech ratings via the AMT platform. The key features of the template were: (a) a collapsible set of instructions providing definitions of comprehensibility, fluency, and accentedness, the three constructs to be rated for each audio file; (b) three 9-point scales for each construct, arranged horizontally, beginning with comprehensibility and ending with accentedness; (c) a short demographic questionnaire for workers; and (d) an optional comment box asking workers to provide feedback on the format and content of the HIT. Audio files were randomized and individually paired with the template, such that each assignment consisted of rating one file, and raters received $0.10 per assignment. Separate HITs for the picture narration and prompted response task were deployed to AMT workers whose IP addresses were located in a Spanish-speaking country. The first batch of ratings was collected over the course of a day, and the second batch remained active for just over a week. In total, four thousand ratings were collected from 55 unique AMT workers, including one non-native speaker.

5.1 Data quality and reliability

Of the 54 native Spanish speakers who participated, 15 were removed from the data set because they either did not complete at least two attention checks (*n* = 12) or did not rate a sufficient number of near-native samples (*n* = 3). Collectively, these 15 workers provided 219 ratings, or 5.48% of the data set. Of the remaining 39 workers, four (10.26%) failed the attention check, responding incorrectly to two or more trials, but none were excluded due to overlapping means for the learner and near-native speaker groups. The fact that approximately 90% of valid workers (i.e., individuals who provided enough ratings to be included in the data set) correctly scored attention checks and consistently rated near-native speakers above intermediate learners suggests that the majority of raters remained attentive throughout the task and accurately discriminated the near-native control files.

Regarding the reliability of AMT data, due to the design of the HIT, the number of clips that each worker evaluated was variable. In other words, each clip was evaluated by a random set of raters drawn from the final pool of 35 valid AMT workers. For this data set, reliability was computed as a one-way, consistency, average-measure ICC. For the sake of comparability with other pronunciation studies, fully-crossed data sets were constructed including only the raters who evaluated all clips (for picture narration, *n* = 18; for prompted response, *n* = 20). For the fully-crossed data sets, reliability was estimated using a two-way, consistency, average-measure ICC. ICC values for comprehensibility and fluency exceeded .90 for both the random raters and fully-crossed data, indicating excellent reliability, and values for accentedness were acceptable but slightly lower. These reliability estimates are similar to those reported in laboratory studies (cf. Table 1), which suggests that crowdsourced data is as reliable as its laboratory counterpart when appropriate measures are taken to ensure that online raters understand and follow instructions.

Rasch analyses generally confirmed this pattern. Although there were significant and reliable differences in rater severity for comprehensibility and fluency, few raters were classified as misfitting, and the corresponding 9-point scales performed as expected along multiple quality dimensions, except category distinctiveness. On the other hand, severity measures for accentedness displayed the largest logit spread, and a larger proportion of raters were classified as misfitting. Furthermore, category frequencies were highly skewed toward the lower end of the scale. These results coincide with Isaacs and Thomson (2013), who reported 11 misfitting raters and category compression for the 9-point scale. The fact that comprehensibility and fluency displayed a much higher degree of internal consistency than accentedness in the current study may be due to rater sampling procedures and the number of near-native samples that each rater evaluated.

Accentedness is typically operationalized with respect to a local variety of the L2, which makes sense for L2 speech studies involving speakers who are living and working (or studying) in a region where a particular variety of the L2 is spoken (e.g., Derwing & Munro, 2013; Kennedy et al., 2015). In this context, local listeners are recruited and instructed to evaluate accentedness in terms of the local variety. This contrasts with the present study, in which listeners from 11 different Spanish-speaking countries (Venezuela, Mexico, and Colombia being the most frequent) were recruited to evaluate classroom language learners who had been exposed to multiple varieties of native and non-native Spanish through their instructors. In this scenario, raters might be expected to agree in terms of their evaluations of comprehensibility and fluency, but diverge somewhat when evaluating accentedness since each individual listener would establish a distinct internal standard guided by the features of the native dialect. Thus, pronunciation features that would be associated with a strong foreign accent in one region might not trigger an equally strong response in another. Even though instructions defined foreign accent in terms of pronunciation features that would not occur in any variety of native Spanish, raters may have struggled to conceptualize accentedness in such broad terms. Moreover, raters may have assumed that individuals learning Spanish in the US would acquire a peninsular variety, since peninsular Spanish is sometimes considered a prestige dialect. Either of these approaches—assessing accentedness with respect to the native dialect or assuming a common, possibly peninsular, target dialect for all learners—could have created instability in the accentedness ratings compared to comprehensibility and fluency, which were assessed more uniformly across dialects.

The design of the task itself may have also created greater range effects for accentedness. In this study, the number of near-native samples that each rater evaluated varied, though all raters included in the final data set evaluated at least two. Given that L2 speakers may be evaluated as more accented when the proportion of native speakers included in the study is higher (Flege & Fletcher, 1992), it could be that AMT raters who evaluated more near-native samples assessed L2 learners as more accented. In all likelihood, all three of these factors—the native dialect of the AMT rater, the presumed target dialect of the learner, and the number of near-native samples rated—contributed to the range of severity values for accentedness and the greater number of misfitting raters.

One final quality metric relates to whether AMT raters detect changes in each dimension of L2 speech as reliably as laboratory raters. Rasch analyses suggest that fluency improved over the course of the study (i.e., over the three data points) and comprehensibility from the first to the last data point (i.e., beginning to end of the academic year). These results generally intersect with the developmental patterns observed in a previous study examining L2 pronunciation development in a sample of novice classroom learners of L2 Spanish (Nagle, 2018a, 2018b). In that study, both comprehensibility and accentedness (fluency was not assessed) improved significantly. Although the present study and Nagle (2018a, 2018b) involve different data sets, the fact that the AMT raters detected changes in comprehensibility and fluency suggests that they are sensitive to changes in L2 speech over time. This aspect of the data should be interpreted with caution, however, until direct comparisons between AMT and lab raters can be made using the same data.

5.2 Rater background

In this study, rater background characteristics did not appear to be related to rater performance in terms of severity/leniency and patterns of misfit or overfit. Isaacs and Thomson (2013) similarly reported that nearly equal numbers of experienced and inexperienced raters—individuals who were ESL teachers or graduate students in other disciplines—were classified as misfitting, though reliability estimates were slightly higher across the board for the experienced group. Taken together, these findings tentatively suggest that certain aspects of rater performance may be idiosyncratic, or at least are related to variables that were not examined in either study. Still more work involving a larger pool of raters is needed before a definitive conclusion can be reached.

In contrast to the rater performance measures, level of education was a significant predictor of the ratings data, in that individuals who had completed a graduate degree evaluated learners' accentedness more positively, assigning them scores nearly one point above the scores they received from raters with a bachelor's or equivalent degree. It could be that this variable actually encodes other experiential factors, if, for example, higher levels of education are correlated with more international experience, more frequent interactions with L2 speakers, or exposure to a wider variety of L1 dialects. This account would be compatible with Bergeron and Trofimovich's (2017) assertion that experience with accented speech may help listeners process it more readily, leading to more positive evaluations.

In broad terms, the demographic characteristics of the workers included in this study indicate that Venezuela, Colombia, Mexico, and Spain are well represented among AMT workers. It might be possible to elicit individual sets of ratings from each of these countries for comparison. At the same time, it is important to bear in mind that the trends reported in this study are indicative of the composition of AMT users at the time of data collection, and that factors related to the task itself may have made it more or less appealing to certain worker subgroups. Thus, more research is needed targeting as wide a sample of L1 Spanish speakers as possible so that more precise data on the number of L1 Spanish users and their countries of origin can be computed.

5.3 Recommendations for improving the AMT task

The findings of this study have implications for improving the collection of L2 speech ratings using crowdsourcing platforms such as AMT. Regarding the workers who were excluded because they failed the attention check, the best solution would be to create a training block that would award raters a special qualification needed to advance to the experimental portion of the ratings. This approach would familiarize workers with the structure and expectations of the HIT, and because AMT enables requesters and workers to leave comments for one another, issues related to instructions, the interface, or the overall concept could be resolved at this stage. This type of screening could be considered equivalent to selecting only high-reputation workers, in which case attention checks might not be necessary (Peer et al., 2014)[4].

A second issue relates to the structure of the task itself. In this study, each assignment consisted of an individual audio file, which meant that raters could complete as many assignments as desired. A one-to-one mapping of audio files to assignments was advantageous because it kept the user interface simple (one set of ratings per screen) and allowed for a larger group of workers to complete assignments simultaneously. Nevertheless, the practical outcome was that some workers who completed a small number of assignments did not evaluate a sufficient number of attention checks or near-native samples. This source of data loss could be resolved by bundling speech samples into batches, such that raters would evaluate sets of 5–10 clips per assignment (e.g., Evanini, Higgins, & Zechner, 2010), and compensation would be increased to match this more complex format. Implementing bundled ratings would allow requesters to set up blocks of files containing one attention check and one near-native sample, the location of each of these quality control files randomized within the block. Including one near-native sample in each block would also combat range effects since the proportion of near-native speakers would be constant across blocks. It would be particularly important to pilot this format since the interface would necessarily be more complex.

Regarding listener demographics, the short demographic survey was embedded directly into the HIT, appearing each time workers initiated a new assignment. This did not seem to confuse workers since instructions made it clear that they only needed to complete the survey once. Nevertheless, if a training block were built, the demographic survey could be incorporated into the screening portion of the experiment, and additional questions regarding familiarity with accented speech and frequency of interaction with non-native speakers could be included.

In addition to the HIT design features discussed above, there are two methodological decisions that merit additional consideration when collecting L2 speech ratings via AMT. In the AMT task employed in this study, raters were allowed to listen to each clip up to three times to prevent momentary technological issues (e.g., issues with the playback device, volume, etc.) from compromising their ability to evaluate the samples. In other words, raters could listen to the clip once per construct, facilitating a sequential ratings paradigm even through all three scales were presented simultaneously. In laboratory studies, raters typically listen to the clip only once and rate all constructs, or listen to the clip once per construct, in which case they rate all files for a given construct before moving on to the next. In this way, files are randomized within each block, preventing repeated exposure from affecting rater intuition. Future research on AMT ratings will need to explore whether or not allowing listeners to replay the clip three times is necessary. If ratings focus on one construct, such as comprehensibility, then one or two opportunities may be sufficient. Collecting more comprehensive data on how AMT users interact with the ratings interface and the particular characteristics of the environment in which they choose to complete ratings (e.g., the time of day, amount of ambient noise, etc.) could shed light on this point.

As noted above, there was greater variability in the accentedness ratings than in either comprehensibility or fluency, which can be attributed to the diversity of L1 dialects sampled and the design of the ratings interface. Bundling samples can help resolve range effects stemming from the number of near-native samples rated, but special care must be taken when operationalizing accentedness for classroom learners and sampling raters to evaluate it. One potential solution would be to integrate students into the rating process by asking them to imagine how they will use the target language in the future. This information could then be used to guide the selection of raters. If students envision themselves studying abroad or living and working in Spain, then raters from Spain could be recruited. Alternatively, if students envision themselves interacting with US Spanish speakers in a particular geographic region, then native and L2 Spanish speakers from that region could be recruited. This strategy could also prove advantageous for other dimensions of L2 speech. For example, research has shown that both lexicogrammatical and pronunciation features contribute to comprehensibility (Saito et al., 2017). Thus, if speakers use words or constructions that do not (frequently) occur in a certain target variety, they may be less comprehensible to that particular group of target interlocutors. Though a more targeted approach to raters is worthwhile, it would be impractical and probably unnecessary to recruit distinct sets of raters for each learner. Instead of adopting a completely individualized approach, listeners could be sampled from a limited number of dialects (e.g., the

2-3 dialects that are most frequently reported as potential future interlocutors), and relationships between L1 dialect and ratings could be modeled as part of hypothesis testing.

6. Conclusion

In this study, a methodology for collecting speech ratings through AMT was presented and validated for L2 Spanish. Results suggest that AMT is an efficient and reliable means of data collection when quality control measures such as attention checks are put into place. While interrater reliability and individual rater metrics revealed a high degree of consensus for comprehensibility and fluency, lower reliability was observed for accentedness. Three primary recommendations were made to improve data collection in AMT: (a) the creation of a training block that would serve as a screening mechanism and special qualification to be awarded to workers; (b) bundling audio files into blocks to ensure that workers complete a minimum number of ratings, including rating attention and control clips; (c) revisiting methodological decisions related to rater sampling and scale length. Given that larger sample sizes and denser longitudinal studies are becoming more frequent, future studies should continue to explore the possibilities that AMT offers for rating linguistic dimensions of speech samples and for mass data collection in other languages. Ultimately, AMT has the potential to advance the state-of-the-art by connecting researchers and teachers working in L2s such as Spanish with a larger and more representative sample of listeners, and researchers and teachers in less commonly taught L2s with listeners to whom they might not otherwise have access.

Notes

1. These two tasks provide a more comprehensive view of speakers’ performance under different task complexity conditions (Crowther, Trofimovich, Isaacs, & Saito, 2015; Crowther, Trofimovich, Saito, & Isaacs, 2018). When describing their routine, speakers

can formulate a response based on their personal experience, avoiding gaps in vocabulary and potentially problematic grammatical structures. In contrast, the picture narrative requires speakers to describe the information included in each frame as accurately as possible and could therefore be characterized as more cognitively demanding.

2. See the supplementary online materials for .csv cross-tabulations summarizing the rater-by-file breakdown for the data set consisting of 54 L1 Spanish workers (i.e., the full data set before trimming based on the control measures).
3. The decision was made to evaluate the unbalanced data set, including as many raters as possible, since this data set subsumes the balanced data set and because crowdsourcing ratings through AMT are likely to lead to data that is not fully crossed, particularly for studies involving a large number of samples.
4. As one reviewer pointed out, the attention checks could confuse raters or interfere with their intuition in scoring the target samples. Although workers were made aware of the fact that they would occasionally receive this type of clip in an effort to avoid any confusion, eliminating attention checks would completely resolve the issue.

Author contact information:

Charles Nagle (cnagle@iastate.edu)
Iowa State University
World Languages and Cultures
3102G Pearson Hall
505 Morrill Road
Ames, IA 50011-2103

References

Akiyama, Y., & Saito, K. (2017). Development of comprehensibility and its linguistic correlates: A longitudinal study of video-mediated telecollaboration. *The Modern Language Journal, 100*(3), 585–609. doi:10.1111/modl.12338

Bergeron, A., & Trofimovich, P. (2017). Linguistic dimensions of accentedness and comprehensibility: Exploring task and listener effects in second language French. *Foreign Language Annals, 50*(3), 547–566. doi:10.1111/flan.12285

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi:10.1177/1745691610393980

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *The Modern Language Journal, 99*(1), 80–95. doi:10.1111/modl.12185

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition, 40*(2), 443–457. doi:10.1017/s027226311700016x

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning, 63*(2), 163–185. doi:10.1111/lang.12000

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197–221. doi:10.1207/s15434311laq0203_2

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

Eskénazi, M., Levow, G.-A., Meng, H., Parent, G., & Suendermann, D. (Eds.). (2013). *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. UK: John Wiley & Sons.

Evanini, K., Higgins, D., & Zechner, K. (2010). *Using Amazon Mechanical Turk for transcription of non-native speech.* Paper presented at the Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, CA.

Flege, J. E., & Fletcher, K. L. (1992). Talker and listener effects on degree of perceived foreign accent. *The Journal of the Acoustical Society of America, 91*(1), 370–389. doi:10.1121/1.402780

Fort, K., Adda, G., & Bretonnel Cohen, K. (2011). Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics, 37*(2), 413–420.

Gelas, H., Teferra Abate, S., Besacier, L., & Pellegrino, F. (2011). Quality assessment of crowdsourcing transcriptions for African languages *Interspeech-2011* (pp. 3065–3068).

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*(3), 213–224. doi:10.1002/bdm.1753

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34.

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. doi:10.1080/15434303.2013.769545

Kennedy, S., Foote, J. A., & Dos Santos Buss, L. K. (2015). Second language speakers at university: Longitudinal development and rater behaviour. *TESOL Quarterly, 49*(1), 199–209. doi:10.1002/tesq.212

Kunath, S. A., & Weinberger, S. H. (2010). The wisdom of the crowd's ear: Speech accent rating and annotation with Amazon Mechanical Turk *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 168–171). Los Angeles, CA: Association for Computational Linguistics.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Martin, D., Hanrahan, B. V., O'Neill, J., & Gupta, N. (2014). *Being a turker*. Paper presented at the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, Baltimore, MD.

McAllister Byun, T., Halpin, P. F., & Szeredi, D. (2015). Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders, 53*, 70–83. doi:10.1016/j.jcomdis.2014.11.003

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. doi:10.1037/1082-989x.1.1.30

Muñoz, C. (Ed.) (2006). *Age and the rate of foreign language learning*. Tonawanda, NY: Multilingual Matters.

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. doi:10.1111/j.1467-1770.1995.tb00963.x

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement, 4*(4), 386–422.

Nagle, C. (2018a). Modeling classroom language learners' comprehensibility and accentedness over time: The case of L2 Spanish. In J. Levis (Ed.), Proceedings of the 9th Pronunciation in Second Language Learning and Teaching Conference (pp. 17–29). Ames, IA: Iowa State University.

Nagle, C. (2018b). 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. doi:10.1111/modl.12461

O'Brien, M. G. (2014). L2 learners’ assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning, 64*(4), 715–748. doi:10.1111/lang.12082

O'Brien, M. G. (2016). Methodological choices in rating speech samples. *Studies in Second Language Acquisition, 38*(3), 587–605. doi:10.1017/s0272263115000418

Paolacci, G., & Chandler, J. (2014). Inside the Turk. *Current Directions in Psychological Science, 23*(3), 184–188. doi:10.1177/0963721414531598

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

Pavlick, E., Post, M., Irvine, A., Kachaev, D., & Callison-Burch, C. (2014). The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics* (Vol. 2, pp. 79–92).

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

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavioral Research Methods, 46*(4), 1023–1031. doi:10.3758/s13428-013-0434-y

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*(3), 709–743. doi:10.1111/lang.12297

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. doi:10.1093/applin/amv047

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition, 15*(4), 905–916. doi:10.1017/S1366728912000168

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.

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement, 14*, 339–355.

Appendix. Worker Characteristics, Files Rated, and Exclusion Criteria.

Worker	Age	Gender	Origin	Education	Files (100)	Checks (14)	Failed Checks	Exclusion Notes
1	32	f	Venezuela	Master	95	13	0	
2	32	f	Venezuela	Bachelor	97	12	2	
3	24	m	Peru	Bachelor	100	14	14	Failed ≥ 2 checks
4	26	m	Spain	Bachelor	100	14	0	
5	35	f	Venezuela	Bachelor	100	14	0	
6	32	m	Venezuela	Master	94	14	0	
7	25	f	Venezuela	Bachelor	82	13	0	
8	31	m	Spain	Bachelor	100	14	0	
9	30	m	Venezuela	Bachelor	97	14	0	
10	26	m	Venezuela	Bachelor	100	14	0	
11	23	m	Venezuela	Bachelor	100	14	0	
12	52	m	Venezuela	Bachelor	100	14	0	
13	36	m	Mexico	Bachelor	100	14	0	
14	48	m	Venezuela	Bachelor	57	8	8	Failed ≥ 2 checks
15	34	m	Venezuela	Bachelor	99	13	3	Failed ≥ 2 checks
16	38	m	Venezuela	Bachelor	32	4	0	
18	35	m	Guatemala	Bachelor	96	14	0	
19	29	f	Spain	Master	36	4	0	
20	21	m	Mexico	Bachelor	71	13	0	
21	30	m	Venezuela	Bachelor	21	4	0	Did not rate ≥ 2 nns files
22		f	Venezuela	Bachelor	30	4	1	
23	24	m	Venezuela	Bachelor	24	3	0	Did not rate ≥ 2 checks
24	43	f	Peru	Bachelor	62	7	0	
25	35	m	Venezuela	Master	19	3	3	Did not rate ≥ 2 checks
26	20	m	Venezuela	HS	31	5	0	Did not rate ≥ 2 nns files
27	30	f	Venezuela	Master	1	0	n/a	Did not rate ≥ 2 checks
28	47	m	Venezuela	Master	16	1	1	Did not rate ≥ 2 checks
29	50	f	Spain	HS	34	3	2	Did not rate ≥ 2 checks
30	43	m	Venezuela	Bachelor	20	4	0	Did not rate ≥ 2 nns files
31	25	m	Venezuela	Bachelor	15	4	4	Failed ≥ 2 checks

32	43	m	Spain	Master	4	0	n/a	Did not rate ≥ 2 checks
33	45	m	Mexico	HS	13	3	1	Did not rate ≥ 2 checks
34	23	f	Costa Rica	Bachelor	94	13	0	
35	45	m	Colombia	Bachelor	100	14	0	
36	33	f	Mexico	Bachelor	100	14	0	
37	25	m	Colombia	HS	65	9	0	
38	22	f	Guatemala	Bachelor	97	14	0	
39	32	m	Colombia	Master	97	13	0	
40	42	m	Mexico	HS	99	14	0	
41	27	m	Colombia	HS	100	14	0	
42	28	f	Colombia	PhD	100	14	0	
43	37	m	Colombia	PhD	100	14	0	
44	24	m	Colombia	Master	53	7	0	
45	36	m	Honduras	Bachelor	89	13	1	
46	48	m	El Salvador	Bachelor	100	14	0	
47	31	m	Mexico	Bachelor	100	14	0	
48	26	m	Colombia	Bachelor	100	14	0	
49	36	m	Mexico	Master	76	11	1	
50	29	m	Argentina	Bachelor	11	1	0	Did not rate ≥ 2 checks
51	25	m	Mexico	Bachelor	16	3	0	Did not rate ≥ 2 checks
52	32	m	Chile	Master	57	10	0	
53	34	f	Mexico	Bachelor	4	1	1	Did not rate ≥ 2 checks
54	33	m	Venezuela	Bachelor	4	0	n/a	Did not rate ≥ 2 checks
55	28	f	Mexico	Bachelor	1	0	n/a	Did not rate ≥ 2 checks

Note. “Files,” “Checks,” and “Failed Checks” refer to the number of files rated, attention checks rated, and attention checks failed. There were 50 files per task, including 39 learner audios, 4 near-native speaker audios, and 7 attention checks. Rater 22 did not report her age. Shaded cells indicate raters that were eliminated because they failed more than two attention checks (*n* = 4), did not rate at least two attention checks (*n* = 12), or did not rate at least two near-native samples (nns; *n* = 3).