Published in final edited form as:

J Commun Disord. 2015 ; 53: 70–83. doi:10.1016/j.jcomdis.2014.11.003.

Online crowdsourcing for efficient rating of speech: A validation study

Tara McAllister Byun, Peter F. Halpin, and Daniel Szeredi

New York University, New York, NY

Abstract

Blinded listener ratings are essential for valid assessment of interventions for speech disorders, but collecting these ratings can be time-intensive and costly. This study evaluated the validity of speech ratings obtained through online crowdsourcing, a potentially more efficient approach. 100 words from children with /r/ misarticulation were electronically presented for binary rating by 35 phonetically trained listeners and 205 naïve listeners recruited through the Amazon Mechanical Turk (AMT) crowdsourcing platform. Bootstrapping was used to compare different-sized samples of AMT listeners against a “gold standard” (mode across all trained listeners) and an “industry standard” (mode across bootstrapped samples of 3 trained listeners). There was strong overall agreement between trained and AMT listeners. The “industry standard” level of performance was matched by bootstrapped samples with $n = 9$ AMT listeners. These results support the hypothesis that valid ratings of speech data can be obtained in an efficient manner through AMT. Researchers in communication disorders could benefit from increased awareness of this method.

Keywords

crowdsourcing; speech rating; research methods; speech sound disorders; speech perception

1. Introduction

To study the efficacy of interventions for speech sound disorder, researchers must identify a valid, reliable method for measuring changes in speech production accuracy or intelligibility over time. Acoustic and other instrumental measures are an indispensable part of this process, but from a clinical standpoint, it is most important to know whether treatment yields a meaningful change in human listeners' perception of speech. Given the potential for bias when raters are familiar with the participants or purpose of a study, it is essential to

© 2014 Elsevier Inc. All rights reserved.

Address correspondence to: Tara McAllister Byun Department of Communicative Sciences and Disorders, New York University 665 Broadway, Room 914 New York, NY 10012, USA Phone: 212-992-9445 Fax: 212-995-4356 tara.byun@nyu.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Learner outcomes: Readers will be able to (a) discuss advantages and disadvantages of data collection through the crowdsourcing platform Amazon Mechanical Turk (AMT), (b) describe the results of a validity study comparing samples of AMT listeners versus phonetically trained listeners in a speech-rating task.

collect these ratings in a blinded, randomized fashion. However, obtaining these ratings can pose a major challenge for speech researchers.

The conventional approach to obtaining ratings of speech data is a multi-step process. First, it is necessary to identify potential raters, typically certified clinicians (e.g., McAllister Byun & Hitchcock, 2012) or students in speech-language pathology (e.g., Maas & Farinella, 2012). Potential raters are generally required to complete some number of practice or training trials, followed by a test to determine their eligibility to participate. If an individual meets a predetermined threshold of accuracy or agreement with a “gold standard” rater, he/she is invited to complete the primary rating task. After ratings are collected, some or all of the responses provided by one listener must be compared against another listener's responses in order to calculate interrater reliability. If agreement between raters falls below a critical threshold, it may be necessary to exclude raters, or to elicit ratings again after providing additional training.

This conventional approach to speech sound rating is well-established in the literature, having produced many examples in which interrater agreement exceeds 80% (Shriberg & Lof, 1991; Shriberg et al., 2010). However, its primary drawback is that it is time-consuming and requires intensive effort on the part of the experimenter as well as the rater. Because most commercially available programs for stimulus presentation and response recording require manual installation of proprietary software, the process typically involves multiple points of in-person contact between the raters and the research team. Researchers may struggle to find an adequate number of listeners, and excluding listeners due to low interrater agreement can represent a major setback. Moreover, obtaining ratings can incur significant costs for the researcher. While it is desirable to use the expert judgment of certified clinicians, many researchers are unable to offer compensation in line with the typical pay rate of speech-language pathologists, who earn a median hourly wage of \$50.00 (American Speech-Language-Hearing Association, 2012). Facing these challenges, researchers may resort to non-optimal methods, such as using the authors or other study personnel as data raters. When listeners are familiar with the experimental design and participants, there is a significant risk that their ratings will reflect some influence of preexisting bias.

A novel solution to the longstanding problem of obtaining speech ratings has arisen in connection with recent technological innovations in the area of online crowdsourcing. In crowdsourcing, a task or problem is assigned not to a small number of specialists, but to a large number of non-experts recruited through online channels. Taken individually, these non-specialists do not display expert levels of performance, but in the aggregate, their responses generally converge with those assigned by experts (Ipeirotis, Provost, Sheng, & Wang, 2014). Crowdsourcing methods have succeeded in solving remarkably complex problems, as in a case where untrained users playing an online game arrived at an accurate model of a protein structure that had previously eluded trained scientists (Khatib et al., 2011). Although computational modeling of related tasks (Ipeirotis et al., 2014) suggests that it should be possible for crowdsourcing to yield speech ratings of comparable accuracy or validity to those assigned by trained listeners, this question has never been investigated empirically.

The present paper investigates this question using the crowdsourcing platform that is best-developed and most widely used at the present time, Amazon Mechanical Turk (AMT). The first section provides an overview of AMT and discusses its advantages and limitations as a tool for research. Subsequent sections describe a specific mechanism developed for online collection of ratings of speech sounds and report the results of a comparison of ratings obtained from trained listeners versus non-specialists recruited through AMT. These findings suggest that crowdsourcing can represent an efficient means of obtaining valid perceptual outcome measures for speech treatment research. For readers interested in adopting AMT for their own research, the appendix offers practical guidelines for setting up data collection on AMT.

1.1 Amazon Mechanical Turk

Amazon Mechanical Turk (AMT) is a crowdsourcing platform where employers can post electronic tasks and members of the AMT community can sign up to complete tasks for pay. Previous work in other disciplines, such as cognitive science and linguistics, has offered in-depth overviews of AMT and guidelines for its use in research. This section summarizes this body of work for the audience of researchers in communication sciences and disorders, drawing in particular on Mason & Suri (2012). The original “Mechanical Turk” was a supposed chess-playing automaton that gained fame in the late 1700s for defeating such formidable opponents as Benjamin Franklin and Napoleon Bonaparte. In the 1820s, it was revealed that the Turk was not a true automaton but was operated by a human chess master hidden inside the machinery. Like the original Mechanical Turk, AMT is characterized as “artificial artificial intelligence”—tasks are performed as if automated, but with human intelligence as the driving force. Amazon originally developed Mechanical Turk as an internal platform for routine tasks that computers do not perform effectively, such as identifying objects in photographs. Now anyone can sign up to post tasks or complete jobs using AMT’s standardized searchable interface. Jobs on AMT are referred to as Human Intelligence Tasks (HITs). Most HITs require only seconds to minutes to complete, and compensation for these microtasks is typically only a few cents. By completing large numbers of HITs, which tend to be simple and repetitive, workers earn an average effective hourly wage that has been estimated at \$4.80 (Ipeirotis, 2010a).

The AMT interface is used by hundreds of thousands of workers, along with roughly 10,000 requestors or employers (Ipeirotis, 2010a; Mason & Suri, 2012). It is possible to complete tasks on AMT from any country, but most workers come from the US and India. Employers can impose geographic restrictions on workers, e.g., allowing only workers with a US-based IP address to view a posted HIT. Based on the self-reported demographics of approximately 3,000 workers across 5 different studies, AMT workers are more likely to be female than male, and they have a median age of roughly 30 years (Suri & Watts, 2011). Most workers do not rely on AMT as their primary source of income; in a survey, nearly 70% of US-based workers selected the statement “fruitful way to spend free time and get some cash” to characterize their motivation for taking jobs on AMT (Ipeirotis, 2010b).

AMT was not designed for the purpose of conducting behavioral research, and most HITs are posted by commercial entities looking to outsource routine tasks such as checking the

accuracy of online product listings. However, recent years have seen an explosion of interest in the possibility of drawing on the AMT community as a vast, inexpensive, and easily accessed pool of research participants. AMT subject pools have been used in published research in cognitive psychology (e.g., Goodman, Cryder, & Cheema, 2013; Paolacci, Chandler, & Ipeirotis, 2010), social psychology (Berinsky, Huber, & Lenz, 2012; Summerville & Chartier, 2013), and clinical psychology (Shapiro, Chandler, & Mueller, 2013). In addition, studies in linguistics have drawn on AMT to obtain grammaticality judgments (Gibson, Piantadosi, & Fedorenko, 2011; Sprouse, 2011), measure phonetic perception (e.g., Kleinschmidt & Jaeger, 2012) and evaluate learning of artificial grammars (e.g., Becker, Nevins, & Levine, 2012; Culbertson & Adger, 2014).

1.2 Advantages of AMT

The primary advantage of using AMT to recruit listeners for speech rating tasks is the speed and ease with which ratings can be obtained. AMT workers constantly monitor the interface for new HITs, and a new task will typically be accepted by multiple workers within minutes of posting. Previous literature has described the rapid rate at which behavioral experiments can be completed via AMT as “revolutionary” (Crump, McDonnell, & Gureckis, 2013). For example, Sprouse (2011) reported the results of a grammaticality judgment task that was completed by 176 native English speakers in a traditional laboratory setting and by the same number of speakers recruited via AMT. While the traditional data collection method required approximately 88 hours of experimenter time, the AMT replication was completed in 2 hours.

AMT can also represent a less expensive alternative to conventional methods of collecting behavioral responses. Typical laboratory subjects tend to cost a bit more than the current minimum wage (Mason & Suri, 2010), and skilled listeners such as speech-language pathologists can command much more than that. Crump et al. (2013) explored different rates of compensation on AMT and reported that many participants were willing to complete a 15-30 minute study when offered only \$0.75. Like previous studies (e.g., Buhrmester, Kwang, & Gosling, 2011), Crump et al. found that the amount of compensation affected the rate at which AMT workers responded to their posting, but it did not have a measurable impact on the quality of the responses received. Overall, it is clear that AMT can offer significant cost savings relative to conventional mechanisms for obtaining ratings of speech data.

Finally, it is important to note that using streamlined methods for speech data rating is not merely a question of convenience; it also has the potential to enhance scientific rigor. As indicated above, the difficulty of obtaining outside raters for speech data can lead to methodological compromises such as using study authors as data raters. By facilitating access to independent raters, AMT could effectively combat this significant source of bias. Second, it is common for researchers to identify limitations of their rating methods over the course of implementation. For example, it may become apparent that instructions are ambiguous, or that the rating scale used was not ideal for the task. In a traditional model, the amount of time and money invested in each round of data collection is so great that any alterations to the protocol must be postponed until the next study. By reducing the resources

expended per experimental cycle, AMT makes it possible to go through multiple iterations of a rating task to achieve the optimal design (Mason & Suri, 2012). Finally, Crump et al. (2013) point out that widespread use of AMT could lead to increased transparency in behavioral research. If researchers report an unexpected or controversial finding, they can opt to make their datasets public, and other researchers can easily run their own replication to support or disconfirm the results reported.

1.3 Limitations of AMT

The greatest intrinsic limitation of AMT is that, due to the remote nature of data collection, there are no direct ways to verify the identity or monitor the progress of individual raters. It is in raters' best economic interest to complete as many HITs per hour as possible, and this can lead to uncooperative behavior such as clicking through at maximum speed with little regard for instructions. In some cases, raters may actually be bots, i.e. computer programs coded to complete HITs on AMT. A level of protection against this behavior is provided by AMT's built-in rating system for workers. Before workers are compensated for a completed HIT, requesters have the option to review the work and reject work that does not meet their standards, and requesters posting a new HIT have the option to accept only workers whose previous work has been accepted at a high rate, typically 90% or greater. With the previous acceptance rate thus affecting their ability to obtain subsequent jobs, workers have an incentive to comply with the instructions of each HIT they attempt. A more pervasive problem is that there is no way to control for the presence of distractions during experimental tasks; users may be watching TV or browsing the web while completing HITs. Thus, raters recruited through AMT may be less focused than raters in the laboratory setting. Subsequent sections will discuss strategies to detect uncooperative raters and also to correct for the increase in noise posed by the potentially lesser degree of attention maintained by AMT participants.

For language-specific tasks, including rating of speech sounds, it is particularly important to narrow the worker pool to native speakers of English (often North American English). As noted above, the requester has the option to screen out international listeners by requiring that workers access the task from a US-based IP address.¹ In addition, requesters can ask screening questions and rule out any participants who do not demonstrate a satisfactory knowledge of English. This topic was discussed in some detail in Sprouse's (2011) study on the use of AMT to collect grammaticality judgments. Sprouse advised against the use of screening tests, because participants can take screening measures multiple times and may therefore pass even without native-level language knowledge. Instead, Sprouse asked participants to answer questions about their language background, and to encourage truthful responding, he made it clear that compensation was not contingent on the response to those questions; data from workers who self-reported non-native speaker status were discarded post hoc. Under these circumstances, only 3 workers self-reported non-native speaker status.

Another limitation that is especially relevant in the context of speech sound rating is the inability to standardize equipment or playback volume in an online rating task. The problem

¹It is possible to falsify an IP address. However, to be paid for AMT tasks, workers must enter bank account information in the Amazon Payments system, which serves as an additional check of their nation of origin.

can be controlled to some extent by asking listeners to follow a standard protocol to prepare for listening (e.g., using headphones, setting system volume to a particular percentage of maximum volume, etc.). Still, this is unquestionably a source of noise in data collection via AMT.

A final challenge is perceived in connection with the need to obtain permission from their institution's Institutional Review Board (IRB) to collect speech rating data via AMT. However, as the use of AMT becomes widespread in psychology and linguistics experiments, IRBs are increasingly familiar and comfortable with proposals incorporating these methods. Because the AMT interface allows research data to be collected in an anonymous fashion, studies conducted exclusively online often fall under the exempt category of research (Schubert, Murteira, Collins, & Lopes, 2013). Even when classified as exempt from review, behavioral studies on AMT often require workers to read and agree to an informed consent document, with the option to proceed to the experiment or return to the main AMT interface.

In the context of speech rating tasks, it is also imperative that the original IRB protocol under which speech samples were collected allow sharing of these samples for the purpose of data rating. In speech treatment research that will involve online rating, the authors use a parental consent form that specifically describes the lab's policy for data sharing and storage. ("We get [accuracy ratings] by playing recordings of participants' utterances in a randomized, de-identified fashion. Below, we will ask for your permission to store recordings of your child's /r/ productions for use in future research. These recordings might be shared with outside listeners in a randomized, de-identified fashion.") In addition, permission to share and store data is not assumed by virtue of the parent's signature on the consent form. Instead, the parent must affirmatively respond to a separate question, "Will you allow us to store audio recordings of your child and share them with outside listeners as part of a future study?" The consent form clearly states that there is no obligation to release audio files for sharing. All methods used in the present study were approved by the Institutional Review Board at New York University.

1.4 Validity of results obtained through AMT

The theoretical rationale for the present application of AMT comes from modeling studies demonstrating that even when individual raters are not highly accurate, aggregating responses across multiple "noisy" raters yields results that generally converge with the ratings assigned by highly accurate experts (Ipeirotis et al., 2014). Because it is not possible to standardize the experimental setup for sound playback or eliminate external sources of distraction over AMT, it is certainly true that an individual AMT listener represents a "noisier" source of rating data than a laboratory-based listener. On the other hand, the rapid and convenient nature of data collection on AMT allows researchers to recruit many more participants than could be obtained through conventional channels. The model developed by Ipeirotis et al. (2014) suggests that with a large enough listener sample, it should be possible to overcome the noise inherent in AMT data collection and arrive at ratings that converge with expert listener judgments.

The validity of data collected through AMT has been taken up as an empirical question in a number of studies in psychology and linguistics. Researchers in behavioral psychology have reported strong agreement between AMT subjects and traditional lab participants in replications of several classic experiments (Crump et al., 2013; Horton, Rand, & Zeckhauser, 2011) or direct comparisons (Paolacci et al., 2010). Likewise, Schnoebelen and Kuperman (2010) reported obtaining similar response patterns from AMT subjects and participants tested in a traditional lab setting for a series of psycholinguistic experiments. Finally, Sprouse (2011) found “no evidence of meaningful differences” (p. 165) between the responses of AMT participants and lab-based subjects in a grammaticality judgment task. On the strength of this evidence, AMT has become a well-established method for behavioral experiments in both psychology and linguistics. However, no published studies to date have used AMT to evaluate data collected in the study of speech development or disorders.

The present study was undertaken as an empirical test of the validity of AMT ratings compared to ratings collected from listeners more representative of the norm in speech research, i.e., clinicians and students in speech-language pathology. The major experimental questions addressed were the following:

Research question 1: Considering a sample of experienced listeners and a sample of AMT listeners, what level of interrater agreement is achieved both within and across groups?

Research question 2: How many AMT listeners need to be sampled in order to attain a level of performance comparable to the standard typically demonstrated by experienced listeners in the published literature?

2. Methods

2.1 Online setup

Tasks on AMT are classified either as “internal HITs,” which are constructed and managed entirely with Amazon's dedicated interface, and “external HITs,” in which users follow a link to a task hosted by a third-party website. Because the standard AMT interface is not optimal for collecting ratings for a large number of sound files, the present study was set up as an external HIT, with participants following a link to the experiment hosted on a separate server. Stimulus presentation and response recording in the present study were accomplished using the open-source experiment presentation software Experigen (Becker & Levine, 2010). Experigen is a web-based framework written in JavaScript, designed primarily for experiments in linguistics. The properties of stimulus presentation are specified in templatic views that can be easily modified by the investigator, while lower level tasks, such as playing sounds or sending results to the server, are handled automatically by the Experigen framework. Experigen provides experimenters with access to a dedicated server for the aggregation of responses entered by users completing the online task. Any files introduced on the client side, such as sound files presented for rating, can be uploaded to an ordinary web hosting service; the present study made use of Simple Storage Service (S3) within Amazon Web Services (AWS). Certain details of this experiment required additional scripting to complement the basic functions of Experigen. In particular, sound management

in Experigen is handled by default with a Flash application that cannot play raw WAV files. To support the WAV file stimuli used in the present study, the Flash program WavPlayer was embedded and integrated with the Experigen core functions. Code for these extensions can be obtained by contacting the authors.

2.2 Stimuli

In the present study, listeners rated 100 single-word utterances containing the English /r/ sound, collected from 15 children at various stages in the process of remediation of /r/ misarticulation. The number of word tokens per child ranged from 2-13, with a mean of 6.7 and a standard deviation of 2.5. These children ranged in age from 6;0 to 15;8 years, with a mean of 9;4 (SD = 28 months). Three of the 15 children were female, consistent with previous findings that speech errors affecting /r/ are more common in male than female speakers (Shriberg, 2009).

All items had previously been scored as correct or incorrect in a randomized, blinded fashion by 3 experienced listeners. These listeners were certified speech-language pathologists who were not familiar with the study or its participants. Additionally, the acoustic properties of the items had been measured. Previous research (e.g., Delattre & Freeman, 1968) has reported that English /r/ can be distinguished from other sonorant phonemes by the low height of the third formant (F3) and a relatively high second formant (F2). This study used F3 – F2 distance in Hertz as the primary acoustic measure, following the finding of Flipsen, Shriberg, Weismer, Karlsson, and McSweeny (2001) that adolescents' correct and incorrect productions of /r/ could most sensitively be characterized using F3 – F2 distance. Measurements were carried out using Praat acoustic software (Boersma & Weenink, 2012). For each token, a trained graduate student visually inspected Praat's automated formant tracking function, adjusted filter order as needed, and identified the minimum value of F3 that did not represent an outlier relative to surrounding points. The first three formants were extracted from a 14-ms Hamming window around this point using Burg's method of calculating LPC coefficients. Ten percent of files were re-measured by a second student rater, yielding an intraclass correlation coefficient of .81. For additional detail on the protocol used to obtain and verify acoustic measurements, see McAllister Byun and Hitchcock (2012). Items were hand-selected so that the 100-word stimulus set featured a roughly normal distribution of F3 – F2 values, as shown in the histogram in Figure 1.² An effort was also made to include approximately equal numbers of items rated correct and incorrect based on the mode across the certified clinician listeners. However, due to the properties of the materials being used (words elicited from children receiving treatment for /r/ misarticulation), there were more items classified as inaccurate ($n = 64$) than accurate ($n = 36$).

²The use of normalized acoustic measures, such as z-scores representing the distance of a given F3-F2 measurement from the mean for the speaker's age, was considered but judged to be non-optimal for the data under consideration here. Normative data for the age range in question are available only for syllabic /ɹ/ (Lee et al., 1999), whereas speech stimuli used in the present study featured /r/ in a range of contexts, including singleton onsets, clusters, and rhotic diphthongs, which have distinct acoustic characteristics (e.g. Flipsen et al., 2001).

2.3 Rater training protocol

As part of a previous study, the authors developed a protocol for online collection of ratings of speech sounds. This protocol has 3 phases: a 20-item training phase, a 100-item eligibility test phase, and the main data-rating phase. Participants are instructed to wear headphones during all study activities. After completing a consent form and a brief demographic questionnaire, participants view a screen of written instructions describing the stimuli and the standards they are expected to apply in rating speech sounds.³ In the training phase, participants hear 20 single-word utterances representing different degrees of accuracy in production of the target sound, drawn from the same sample of speakers used for the 100-word stimulus set described above. During these and all subsequent trials, participants also see the target word in standard orthography, and they can click to replay a sound file up to 3 times. Participants are instructed to rate the target sound in each word by clicking “Correct” or “Incorrect,” and they receive immediate feedback regarding the accuracy of their response. For this purpose, accuracy can be determined based on ratings previously assigned by experienced listeners; the present study made use of the mode across blinded ratings obtained from 3 certified clinicians. After completing 20 trials, participants receive a score representing their percent agreement with the experienced listener ratings, and they are given a choice to discontinue, practice again, or move on to the eligibility phase. In the eligibility phase, participants hear 100 single-word utterances containing the target sound, presented in random order without feedback. The accuracy of participants’ responses to these items can be evaluated relative to a standard that draws on acoustic measurements and/or expert listener ratings, as described below. In the protocol developed for the authors’ previous study, participants who passed the eligibility criterion were advanced to a main experimental rating phase in which they heard isolated words containing the target sound in blocks of 200 items. For the purpose of the present validation study, testing was complete after the block of 100 trials corresponding with the eligibility test phase.

Two measures were built into the experimental protocol to ensure that participants were performing the task adequately. First, to identify participants who might not be sustaining an adequate level of attention to the task, each block contains 20 catch trials that were judged by the experimenters to be unambiguously correct or incorrect. If a participant did not score above chance (determined to be 16/20 based on a 2-sided binomial test with $\alpha = .01$) on the catch trials in a block, data from that participant were discarded. Second, a criterion that combines acoustic measures and experienced listener ratings was developed to evaluate raters’ performance and discard unsatisfactory raters. The nature of the materials and criteria used to include/exclude raters will differ across studies; for the present study rating children’s /r/ sounds, it was required that the mean F3 - F2 distance of /r/ sounds rated incorrect exceed the F3 - F2 distance of the /r/ sounds rated correct by a predetermined minimum threshold. This threshold was determined based on 3 certified clinicians’ blinded ratings of the same stimuli. For each clinician, the distance between the mean F3 – F2 value

³For example, the text of the introductory screen reads “You will listen to a series of words containing the English “r” sound. These words were produced by English-speaking children of different ages. Some of the “r” sounds are correct, and some are incorrect. Your task is to rate the accuracy of the “r” sound in each word by clicking on the appropriate label: {correct} {incorrect (any error or distortion)}. We want you to use a strict standard, so that only true, adult-like “r” sounds receive the correct rating. You do NOT need to assign equal numbers of correct and incorrect ratings.

for /r/ sounds rated incorrect and those rated correct was calculated, and this value was averaged across the 3 experienced listeners. The adequacy of a new listener's ratings was tested by calculating a 95% confidence interval around the difference in Hertz between the set of stimuli rated correct and the set rated incorrect for that listener. If a listener's confidence interval included the experienced raters' mean distance and did not include zero, his/her data were retained for analysis; otherwise, the data were discarded.

2.4 Participants

AMT was used to recruit 205 self-reported native English speakers with US-based IP addresses. Participants received \$0.75 for completing the 20-word training and the 100-word sample. On AMT, data collection was completed in 23 hours at a cost of \$167, including Amazon fees. Of 205 listeners, 33 were excluded due to chance-level performance on catch trials, 14 were excluded for failure to meet the acoustic eligibility criterion, and 5 were excluded for completing less than the full set of 100 trials. The final sample consisted of 153 listeners with a mean age of 33.4 years (SD 11.0 years; range 19-61 years). None reported a history of speech or hearing impairment.

To represent experienced listeners, a second sample was recruited that was composed of speech-language pathologists, linguists, and students in linguistics or communication disorders. These participants were recruited through postings to listservs and social media channels targeting these communities, as well as announcements at professional conferences and continuing education events. Direct compensation was not offered, but participants were entered into a drawing for a \$25 gift card. Data collection from experienced listeners was completed over a period of approximately 3 months. Complete responses were obtained from 35 participants. Results were discarded from 9 participants who described themselves as native speakers of a dialect other than American English, as well as one listener who failed to pass the attentional catch trials. All other experienced listeners scored above chance on the catch trials and passed the acoustic eligibility criterion, resulting in a total of 25 listeners with a mean age of 32.1 years (SD 8.9 years; range 21-57 years). Of these respondents, 4 reported completing a bachelor's degree or some coursework in speech-language pathology or linguistics; 20 reported completing a master's degree, and 1 reported completing a PhD in one of these fields.

2.5 Analyses

2.5.1 Research question 1: Group-level analyses—To address the first experimental question, rating behavior was compared across the full set of listeners in each category, i.e. all AMT listeners versus all experienced listeners. Items were classified to reflect agreement or lack of agreement between the modal rating across the full set of AMT listeners and the modal rating across the full set of experienced listeners. Items were assigned to the “Agree (Correct)” category if they were classified as correct by at least 50% of AMT listeners and at least 50% of experienced listeners, and to the “Agree (Incorrect)” category if they were rated correct by <50% of listeners in each group. A third category, “Disagree,” was established for cases where the mode across AMT listeners differed from the mode across experienced listeners. The correlation between the percentage of experienced listeners and the percentage of AMT listeners who rated a given item as correct was calculated. In addition, the

correlation between F3 – F2 distance and percentage of individuals assigning the “Correct” rating was calculated across items for both experienced and AMT listeners.

2.5.2 Research question 2: Bootstrap analyses—While the group-level analyses address the overall agreement between experienced and naïve listeners, it is also desirable to know how many AMT listeners are needed to attain a level of performance comparable to the standard typically achieved by experienced listeners in the published literature. This question was addressed using bootstrap sampling, in which repeated random sampling with replacement is used to estimate the properties of the underlying population from which the sample data were drawn (Efron, 1979). The bootstrap analysis examined the properties of different-sized samples of naïve listeners, ranging from a minimum n of 3 to a maximum n of 15. The value of 15 was taken as the upper limit because previous research has found that gains in rating accuracy associated with increasing numbers of naïve raters reach an asymptote before this point (Ipeirotis et al., 2014). Only odd-numbered sample sizes were used so that modal rating could be calculated without ties. For each value of n , the bootstrap algorithm draws $N = 1000$ random samples of AMT raters with replacement. For each sample drawn, the modal rating across the selected n raters is computed for each stimulus item. For each item and each value of n , the algorithm outputs the proportion of sampling runs in which the modal rating agreed with a predetermined gold standard. The gold standard used as the point of comparison in the present analysis was the modal rating across the full set of 25 experienced listeners. To aggregate over items, results are also reported in terms of the proportion of items for which a threshold of 80% agreement with the gold standard was reached, at each value of n . The target level of 80% agreement with the gold standard was selected to reflect typical expectations for published research in speech-language pathology. Finally, the proportion of items reaching the 80% agreement threshold for different-sized samples of AMT listeners was compared against an “industry standard” that is described in detail below.

At this point, three details of the bootstrap analysis require further explanation. The first is the decision to use the modal rating as the method of aggregating scores across raters. The mode yields a binary rating and is referred to as “hard” classification; it can be contrasted with methods of “soft” classification, such as the mean over raters, which represent the probability that an item will be rated correct. Previous research on crowdsourcing has reported comparable results from hard and soft classification (Ipeirotis et al., 2014). Because many clinical applications, such as scoring standardized tests, require a binary correct/incorrect outcome, the “hard” classification method (modal rating) was used throughout the analyses reported here. Second, the decision to adopt the mode across all 25 experienced listeners as the gold standard reflects the focus of the present study on comparing the performance of a new source of ratings (i.e., crowdsourcing) to that of an established standard (i.e., experienced listeners). Naturally, other gold standards could be used to address different research questions.

The third issue requiring further explication is the decision to compare AMT ratings not only against a gold standard, but also against a more representative “industry standard.” Few studies in the literature have made use of a pool of experienced raters as large as the 25-listener sample that formed the basis for the gold standard adopted here. Therefore, the

“industry standard” was operationalized to represent the norm for published studies reporting speech rating data. While there are no formal guidelines, a minimum of 3 raters achieving at least 80% agreement with each other or with an external gold standard might reasonably be regarded as the basic expectation for published research. Therefore, the “industry standard” was estimated by conducting an additional bootstrap analysis in which $N = 1000$ sets of $n = 3$ experienced listeners were randomly selected with replacement from the larger pool of 25 experienced listeners. The percentage of sampling runs in which the mode across 3 listeners agreed with the gold standard was calculated for each item, as well as the proportion of items in which agreement with the gold standard reached 80%. Finally, as described in the results section, the experienced raters did not attain a high level of agreement for all 100 items. In particular, 10 items that had acoustically intermediate characteristics (near the middle of the range of F3 - F2 values) showed less than 60% agreement across all experienced listeners. These items also failed to reach 80% agreement with the gold standard in the bootstrap analysis across experienced raters, even with sample sizes up to $n = 15$. These items were characterized as inherently ambiguous; they appear to fall in the perceptual boundary region between /r/ and another phoneme. The mode across 25 experienced listeners does not appear to represent a valid gold standard for these items, since large samples of experienced listeners fail to demonstrate an adequate level of agreement with their own modal rating. These items were thus omitted from the bootstrap analysis used to evaluate AMT listener accuracy. The question of how to obtain valid ratings for ambiguous items like these will be addressed in the discussion section.

3. Results

3.1 Research question 1: Group-level analyses

Figure 2 shows the percentage of AMT listeners who assigned the “Correct” rating to a given item, plotted as a function of the percentage of experienced listeners who rated the same item as correct. Stimulus items in Figure 2 are differentiated to reflect the above-described categories “Agree (Correct),” “Agree (Incorrect),” and “Disagree.” Out of 100 items, 37 fell in the “Agree (Correct)” category, 56 fell in the “Agree (Incorrect)” category, and only 7 fell in the “Disagree” category. In 6/7 cases, the percentage of AMT listeners assigning the “Correct” rating was higher than the percentage of experienced listeners scoring the item as correct, suggesting that AMT listeners applied a somewhat more lenient standard than experienced listeners when rating the present speech stimuli. Across items, there was a large and highly significant correlation between the percentage of experienced listeners and the percentage of AMT listeners who rated a given item as correct ($r = .98, p < .001$). Thus, at the level of the full listener sample, strong agreement was observed between AMT listeners and experienced raters.

The 100 speech stimuli yielded differing percentages of agreement across raters: some items were virtually 100% agreed-upon by both experienced and AMT listeners, while others showed chance-level rating behavior even among experienced listeners. For both groups, the percentage of agreement across raters differed as a function of the acoustic characteristics of the stimulus item, as seen in Figure 3. Items with very small (accurate) or very large (inaccurate) F3-F2 distances were generally associated with high levels of interrater

agreement for both experienced and AMT listeners, while those with intermediate acoustic properties yielded lower levels of agreement. Most notably, the 7 points in the “Disagree” category were associated with F3 – F2 distances ranging from 947 Hz – 1632 Hz, falling near the middle of the range for the sample as a whole (40 Hz – 2563 Hz). The correlation between F3 – F2 distance and percentage of “Correct” ratings was large and highly significant for both experienced listeners ($r = -.77, p < .001$) and AMT listeners ($r = -.79, p < .001$).

3.2 Research question 2: Bootstrap analyses

Figure 4 depicts the results of the bootstrap analysis over AMT listener samples. As discussed above, only the 90 items that showed at least 60% agreement across the full sample of experienced listeners were included in this analysis. Each item is represented by a series of potentially overlapping circles of differing sizes. The size of the circle represents the number of AMT raters included in the randomly drawn sample in each iteration of the bootstrap analysis; odd-valued sample sizes ranging from $n = 3$ to $n = 15$ are represented for each item. The x-axis represents the acoustic properties (F3 – F2 distance in Hertz) of each item. The y-axis represents the percentage of sampling runs in which the mode across listeners in the bootstrap sample agreed with the gold standard rating for a given stimulus item. Finally, a horizontal dotted line marks the target threshold representing agreement with the experienced listener gold standard in 80% of sampling runs.

Several properties stand out from this visual depiction of the bootstrap results. First, as a general rule, as the number of AMT listeners included in the randomly selected sample increases, percent agreement with the gold standard also increases. There are only 3 items for which the AMT listeners truly diverged from the experienced listeners, such that increasing the number of AMT listeners included in the randomly selected sample results in a lower level of agreement with the gold standard. Second, there is extensive item-by-item variability in the maximum level of agreement attained. Consistent with the findings reported in the previous section, items with intermediate F3 – F2 values tended to be associated with lower maximum levels of agreement than items with extreme acoustic properties.

As discussed above, a decision was made to summarize over items using the percentage of stimulus items for which the AMT listener sample matched the gold standard in at least 80% of sampling runs. This percentage was calculated using only the 90 items for which the full sample of experienced listeners showed at least 60% agreement. Figure 5 plots this value for all odd-numbered sample sizes ranging from a minimum of 3 to a maximum of 15 AMT listeners. The result for the “industry standard,” a bootstrap sample over experienced listener data with $n = 3$, is also plotted on the same axes. Figure 5 shows increasing agreement as the number of AMT listeners in each sample increases from 3 to 9. With 3 AMT listeners, the 80% agreement threshold is reached by 69/90 stimulus items (77%). With 5 listeners, the 80% threshold is reached by 73/90 items (80%), and with 9 listeners, it is reached by 77/90 items (86%). The “industry standard” of bootstrap samples of 3 experienced listeners demonstrated exactly the same performance (77/90 items, 86%). After the sample size $n = 9$, a plateau can be observed in the performance of samples of AMT listeners; there is no

further improvement in the percentage of items reaching the 80% agreement threshold for sample sizes up to 15 listeners.

4. Discussion

The study's first experimental question pertained to the overall level of agreement between experienced listeners and AMT listeners when responses were aggregated across all individuals in a sample. Results indicated a very high level of overall agreement, with a correlation of .98 between the percentage of experienced and AMT listeners rating a given item as correct. The second question asked how many AMT raters would be necessary to achieve a level of agreement with the gold standard that is consistent with typical expectations for published speech rating data. The present finding was that samples of 9 or more AMT listeners demonstrate a level of performance equivalent to that of the “industry standard” adopted to represent the norm of rater performance typically upheld in the published literature.

Inspection of individual stimulus items revealed that the level of agreement attained, both within and across groups, was strongly predicted by the acoustic properties of an item. All 7 points of disagreement between the AMT and experienced listener modes were associated with intermediate acoustic characteristics. It is of interest to consider how these across-stimulus differences bear on the question of how many AMT listeners are needed to achieve an adequate level of agreement with the experienced rater gold standard. In their investigation of the properties of ratings drawn from large numbers of non-expert raters, Ipeirotis et al. (2014) reported that increasing the number of raters yielded improvements in rating performance only for items within a specific band of rating difficulty. For very easy items, maximum accuracy was reached with only a small number of raters, while for very difficult items, raters were guessing randomly, and accuracy was not improved by the inclusion of additional random guesses. This finding was replicated in the present study. Acoustically extreme items reached or exceeded 80% agreement with the gold standard for even the smallest sample size of AMT raters, while a number of acoustically intermediate productions did not achieve 80% agreement even at the largest sample size—and this low level of agreement was observed for ratings aggregated across experienced as well as AMT listeners.

In the present study, 10 items that were not classified reliably even by experienced listeners were omitted from the bootstrap analysis used to compare AMT listeners against experienced listeners. However, any task of rating speech samples from children and/or individuals with disordered speech will inevitably involve items that are acoustically and perceptually intermediate. One option is to offer a greater number of rating options, such as “fully correct,” “intermediate/distorted,” and “fully incorrect” categories (e.g., Klein, Grigos, McAllister Byun, & Davidson, 2012). However, previous research suggests that increasing the number of categories on the scale in a speech-rating task tends to make point-to-point interrater agreement worse rather than better; McAllister Byun & Hitchcock (2012) described “particular lack of agreement as to which error sounds should be considered distortions versus true substitutions” (p. 212). Work by Munson and colleagues (e.g. Munson, Schellinger, & Urberg-Carlson, 2012) has offered evidence that accurate

perceptual measures of gradient degrees of difference speech can be obtained through the use of Visual Analog Scaling (VAS). However, Payesteh & Munson (2013) found poor interrater agreement when visual analog scales were used for children's /r/ sounds. More fundamentally, there is inherent tension between the use of Visual Analog Scaling and the notion that speech perception is categorical (i.e., while listeners are not insensitive to within-category distinctions, they detect them less reliably than boundary-crossing phonemic contrasts). An ideal solution would capture the gradience of child speech without requiring listeners to override this natural tendency to perceive discrete phonemic categories. Such a solution is available through Item Response Theory (IRT), which models the probability of a given response as a function of both item and person/rater characteristics (e.g., Baylor et al., 2011). When many listeners categorize items in a binary fashion, individual biases yield different boundary locations, and IRT can be used to derive a continuous measure of category prototypicality for each item. In the past, this was not a practical form of measurement because of the slow and labor-intensive nature of collecting speech data ratings. This obstacle is eliminated entirely when a crowdsourcing platform like AMT is used to obtain listener ratings. Follow-up research, now in preparation, will compare gradient ratings derived via IRT from AMT listener judgments against acoustic measurements of the same items.

Lastly, it is important to consider that for some purposes, ratings assigned by naïve listeners on AMT may be preferable to judgments rendered by expert clinicians. Intervention for speech disorders is provided with the goal of making a change in a speaker's functional communication abilities. If a change is so fine-grained that it is detected only by phonetically trained listeners, its functional impact is likely to be limited. By presenting speech samples to naïve listeners on AMT, a researcher can easily evaluate whether a treatment has had a significant impact on intelligibility or perceived accuracy relative to the types of listeners with whom an individual affected by speech disorder is likely to interact in his/her daily life. To date, few interventions have been evaluated in terms of their efficacy in changing treated participants' accuracy or intelligibility to naïve listeners. AMT provides an extremely simple, straightforward means of obtaining such data. Adoption of AMT among researchers in speech-language pathology could thus have a substantial impact on the process by which the efficacy of speech intervention methods is evaluated.

5. Conclusions

This paper aimed to introduce Amazon Mechanical Turk to the audience of researchers in communication sciences and disorders, with a focus on the application of AMT to obtain ratings of speech samples. Like any experimental method, data collection via AMT has both benefits and limitations. Compared to conventional methods of obtaining ratings of speech stimuli, AMT offers undeniable advantages in speed and convenience. However, this convenience comes at the cost of reduced control over the experimental setup, with the consequence that researchers must take careful steps to identify appropriate raters and to discard data suggestive of inattentive or uncooperative rater behavior. Theoretical justification for using AMT in spite of this loosening in experimental control comes from previous research demonstrating that when responses are aggregated across multiple individuals, even low-quality, "noisy" raters can return results that converge with expert

ratings. The validation study reported here indicated that when responses are pooled across samples of at least 9 raters, naïve listeners recruited through AMT demonstrate a level of performance consistent with expectations typically upheld in the published literature. Since AMT samples of this size can be obtained extremely quickly, easily, and cheaply, many speech researchers could find it beneficial to make use of AMT in place of traditional methods of collecting ratings from experienced listeners.

It could also increase experimental rigor by making it easier for researchers to obtain ratings from independent, blinded listeners instead of making use of study personnel, whose familiarity with the experimental design and/or the speakers under evaluation constitutes a source of bias. At the present time, AMT has become a well-established tool in psychology and linguistics, and it has had a profound impact on the way behavioral research is conducted in those fields. In light of the favorable findings reported in the present study, investigators in communication sciences and disorders can be encouraged to explore the potential of this tool for their own research.

Acknowledgments

The authors extend their gratitude to Michael Becker for assistance with the Experigen platform, to Mario Svirsky for helpful commentary, and to all of the students and speech-language pathologists who completed our listening task. Aspects of this research were presented at the Motor Speech Conference in Sarasota, FL (2014) and the Workshop on Crowdsourcing and Online Behavioral Experiments in Palo Alto, CA (2014). This project was supported by an NIDCD R03 to the first author.

Appendix. Practical guidelines for obtaining speech ratings through AMT

1. Getting started

Anyone can sign up to post a job on AMT by creating an account at requester.mturk.com. Since virtually all HITs offer monetary compensation, most requesters also create an account at Amazon Payments, which requires a US credit card or bank account. To obtain ratings of speech sounds, it is also necessary to have an online location to host sound files. For ease of integration with AMT, the present study used Amazon Web Services' (AWS) Simple Storage Service (S3).

Requesters are not required to use their real names in their postings, but it is appropriate to select a username that clearly identifies the source of the HIT (e.g., "NYU Child Speech Lab"). When creating an AMT requester account, it is worthwhile to keep in mind that the account will be the researcher's face to the AMT worker population, with whom it is valuable to establish a good rapport. AMT workers maintain an active online community where they discuss raters and HITs. If a requester obtains a reputation for poor compensation or unwarranted rejections, he/she may find it difficult to engage AMT workers in future studies. It is good practice for researchers to sign up for turkernation.com; perusing the message boards will give the requester a sense of accepted practices and will also make it possible to monitor one's own reputation in the community. There is a dedicated message board where new requesters can introduce themselves to the community and solicit feedback on HITs.

2. Posting a HIT

To post a HIT on AMT, the requester must create a task and also specify a number of supplementary pieces of information. Jobs posted on AMT can be categorized as internal or external HITs. An internal HIT is constructed using Amazon's dedicated interface, while an external HIT consists of a link to a third-party website. Speech sound rating is most easily accomplished using an external HIT. The following section describes how to use the Experigen server to present sound files and record ratings assigned by listeners recruited through AMT.

When posting a HIT, the requester must supply a title and a description that will allow AMT workers to decide whether or not they are interested in accepting the HIT. The requester must then provide additional information, including the compensation offered, the number of unique workers desired to complete each HIT, the time allotted per HIT, and the time the HIT will expire. Setting compensation for a HIT can be a matter of trial and error; Crump et al. (2013) suggest starting at a rate somewhat below the estimated reservation wage and increasing the amount if HITs are not being accepted quickly enough. It is also possible to offer bonus compensation to workers. Thus, one way to increase the attractiveness of a job is to add a lottery component where one or more workers are randomly selected to receive a larger amount of compensation in the form of a bonus. To customize requirements for workers (i.e. who will be eligible to complete the task), it is necessary to click the advanced options link at the bottom of the AMT requestor interface. It is typical to require that previous jobs completed by a worker have been accepted at a rate of 90% or better. For speech sound rating, it will typically be appropriate to restrict workers by country of origin, accepting only responses from US-based IP addresses. Before making a HIT available to the AMT worker community, it is possible to preview its appearance and test its functioning using the Mechanical Turk requestor "sandbox" (<https://requestersandbox.mturk.com/>).

A final step in launching an online experiment is to specify the set of files or links associated with the HIT. HITs typically are presented as batches of related tasks that have the same directions and just use different files or images. For example, in the present study, each HIT was a block of 200 sound files, and HITs were uploaded in batches of 10 blocks. It is important to note that AMT workers tend to search for and favor tasks that have large batches of HITs available: because it takes time to read through directions and become familiar with a task, workers can maximize efficiency by completing multiple HITs in a single batch.

3. Managing HITs in Experigen

For external HITs, the AMT standard is to issue each worker a randomly generated code upon completion of the task. To manage payment to participants, the study script includes a step generating and presenting a unique code following completion of all experimental stimuli. The participant is directed to return to the AMT interface and enter this code in order to receive credit for his/her work. When the worker enters this code in the AMT interface, the task is marked as completed, and it is passed to the requester for review. This review process provides the requester with an opportunity to filter out poor-quality workers

or bots. The requester could look for responses that are too fast to be genuine, or for suspicious patterns such as providing the same response over and over (Zhu & Carterette, 2010). Another option is to include some items with verifiable answers (for example, speech sounds that are unambiguously correct or incorrect), track a worker's accuracy on those items, and reject work that falls below a predetermined accuracy threshold. However, since rejecting work affects AMT workers' ability to obtain other HITs, it is generally appropriate to limit rejections to cases where the worker was unambiguously engaged in uncooperative behavior. The instructions should also indicate how workers' performance will be evaluated, and if work will be rejected based on incorrect responses to pre-scored questions, this should be stated clearly. Lastly, web browser cookies can be used to track and limit the number of experimental blocks done by one subject.

References

- American Speech-Language-Hearing Association. 2012 schools survey report: SLP workforce/work conditions. 2012 www.asha.org/research/memberdata/schoolsurvey/.
- Baylor CR, Hula W, Donovan NJ, Doyle P, Kendall D, Yorkston KM. An introduction to Item Response Theory and Rasch models for speech-language pathologists. *American Journal of Speech-Language Pathology*. 2011; 20:243–259. [PubMed: 21622595]
- Becker, M.; Levine, J. Experigen - an online experiment platform. 2010. Retrieved from <https://github.com/tlozoot/experigen>
- Becker M, Nevins A, Levine J. Asymmetries in generalizing alternations to and from initial syllables. *Language*. 2012; 88(2):231–268.
- Berinsky AJ, Huber GA, Lenz GS. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*. 2012; 20(3):351–368.
- Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*. 2011; 6(1):3–5.
- Crump MJ, McDonnell JV, Gureckis TM. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One*. 2013; 8(3):e57410. [PubMed: 23516406]
- Culbertson J, Adger D. Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*. 2014 doi: 10.1073/pnas.1320525111.
- Delattre P, Freeman DC. A dialect study of American r's by X-ray motion picture. *Linguistics*. 1968; 6(44):29–68.
- Efron B. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*. 1979; 7:1–26.
- Flipsen P, Shriberg LD, Weismer G, Karlsson HB, McSweeney JL. Acoustic phenotypes for speech-genetics studies: Reference data for residual /3/ distortions. *Clinical Linguistics & Phonetics*. 2001; 15(8):603–630.
- Gibson E, Piantadosi S, Fedorenko K. Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*. 2011; 5(8):509–524.
- Goodman JK, Cryder CE, Cheema A. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*. 2013; 26(3):213–224.
- Horton JJ, Rand DG, Zeckhauser RJ. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*. 2011; 14(3):399–425.
- Ipeirotis PG. Analyzing the Amazon Mechanical Turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*. 2010a; 17(2):16–21.
- Ipeirotis, PG. Demographics of Mechanical Turk. Vol. 10. Center for Digital Economy Research Working Papers; 2010b. Available at <http://hdl.handle.net/2451/29585>
- Ipeirotis PG, Provost F, Sheng VS, Wang J. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*. 2014; 28(2):402–441.
- Julien HM, Munson B. Modifying speech to children based on their perceived phonetic accuracy. *Journal of Speech Language and Hearing Research*. 2012; 55(6):1836–1849.

- Khatib F, Cooper S, Tyka MD, Xu K, Makedon I, Popovi Z, Players F. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*. 2011
- Klein HB, Grigos MI, McAllister Byun T, Davidson L. The relationship between inexperienced listeners' perceptions and acoustic correlates of children's /r/ productions. *Clinical Linguistics and Phonetics*. 2012; 26(7):628–645. [PubMed: 22690718]
- Kleinschmidt, DF.; Jaeger, TF. A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation.. Paper presented at the 34th Annual Conference of the Cognitive Science Society; Sapporo, Japan. 2012.
- Lee S, Potamianos A, Narayanan S. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America*. 1999; 105(3): 1455–1468. [PubMed: 10089598]
- Maas E, Farinella KA. Random versus blocked practice in treatment for childhood apraxia of speech. *Journal of Speech, Language and Hearing Research*. 2012; 55(2):561–578.
- Mason W, Suri S. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*. 2012; 44(1):1–23. [PubMed: 21717266]
- McAllister Byun T, Hitchcock ER. Investigating the use of traditional and spectral biofeedback approaches to intervention for /r/ misarticulation. *American Journal of Speech-Language Pathology*. 2012; 21(3):207–221. [PubMed: 22442281]
- Munson B, Schellinger SK, Urberg-Carlson K. Measuring speech-sound learning using visual analog scaling. *Perspectives in Language Learning and Education*. 2012; 19:19–30.
- Paolacci G, Chandler J, Ipeirotis PG. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*. 2010; 5:411–419.
- Payesteh, B.; Munson, B. Clinical feasibility of Visual Analog Scaling.. Poster presented at the 2013 Convention of the American Speech-Language Hearing Association; Chicago, IL.. 2013.
- Schnoebelen T, Kuperman V. Using Amazon Mechanical Turk for linguistic research. *Psihologija*. 2010; 43(4):441–464.
- Schubert TW, Murteira C, Collins EC, Lopes D. ScriptingRT: A software library for collecting response latencies in online studies of cognition. *PLoS ONE*. 2013; 8(6):e67769. [PubMed: 23805326]
- Shapiro DN, Chandler J, Mueller PA. Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*. 2013; 1:213–220.
- Shriberg, LD. Childhood speech sound disorders: From post-behaviorism to the postgenomic era.. In: Paul, R.; Flipsen, PJ., editors. *Speech sound disorders in children: In honor of Lawrence D. Shriberg*. Plural Publishing; San Diego, CA: 2009. p. 1-34.
- Shriberg LD, Lof GL. Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics & Phonetics*. 1991; 5:225–279.
- Shriberg L, Fourakis M, Hall SD, Karlsson HB, Lohmeier HL, McSweeney JL, Wilson DL. Perceptual and acoustic reliability estimates for the Speech Disorders Classification System (SDCS). *Clinical Linguistics and Phonetics*. 2010b; 24(10):825–846. [PubMed: 20831379]
- Sprouse J. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*. 2011; 43(1):155–167. [PubMed: 21287108]
- Summerville A, Chartier CR. Pseudo-dyadic “interaction” on Amazon's Mechanical Turk. *Behavior Research Methods*. 2013; 45(1):116–124. [PubMed: 23055159]
- Suri S, Watts DJ. Cooperation and contagion in web-based, networked public goods experiments. *PLoS One*. 2011; 6(3):e16836. [PubMed: 21412431]
- Wolfe V, Martin D, Borton T, Youngblood HC. The effect of clinical experience on cue trading for the /r-w/ contrast. *American Journal of Speech-Language Pathology*. 2003; 12(2):221–228. [PubMed: 12828535]
- Zhu, D.; Carterette, B. An analysis of assessor behavior in crowdsourced preference judgments.. Paper presented at the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation; Geneva, Switzerland. 2010.

Highlights

- Obtaining ratings of speech data in the laboratory setting is a slow and effortful process.
- Individual raters obtained through crowdsourcing platforms like Amazon Mechanical Turk (AMT) are “noisier” than laboratory raters, but aggregating across multiple noisy raters can yield convergence with expert ratings.
- Bootstrap sampling was used to compare different-sized samples of AMT listeners against a “gold standard” and an “industry standard.”
- Samples of 9 or more AMT listeners converged on the same rating behavior exhibited by the “industry standard.”
- Crowdsourcing is a well-established tool in linguistics and psychology, and the field of communication disorders could benefit from broader adoption of this method.

FIGURE 1.

Distribution of stimulus items across F3 – F2 values and perceptually correct/incorrect rating categories. Classification as correct/incorrect reflects the mode across binary ratings assigned by three blinded certified clinicians.

FIGURE 2.

Percentage of experienced listeners and AMT listeners rating a given stimulus item as correct

FIGURE 3.

Percentage of experienced listeners and AMT listeners rating a given stimulus item as correct, as a function of F3 – F2 distance. Shaded band represents a 95% confidence interval around the best-fit line.

FIGURE 4.

Percentage of sampling runs in which the mode across bootstrapped AMT listener samples matched the gold standard rating. Size of circle represents number of listeners included in sample, with n ranging from 3 to 15.

FIGURE 5.
Percent agreement with gold standard as a function of size of AMT listener sample in bootstrap analysis