Munro, M. J., & Derwing, T. M. (2020). Collecting data in L2 pronunciation research. In O. Kang, S. Staples, K. Yaw, & K. Hirschi (Eds.), *Proceedings of the 11th Pronunciation in Second Language Learning and Teaching conference*, ISSN 2380-9566, Northern Arizona University, September 2019 (pp. 8–18). Ames, IA: Iowa State University.

COLLECTING DATA IN L2 PRONUNCIATION RESEARCH

Murray J. Munro, Simon Fraser University
Tracey M. Derwing, University of Alberta; Simon Fraser University

Obtaining high quality data in L2 pronunciation research requires careful attention to details at multiple levels. In this paper we share our recommendations on data collection by reflecting on our experiences in a number of studies. We limit our focus to the research designs and data types that we ourselves are familiar with. In particular, we explore approaches to measuring the constructs of intelligibility, comprehensibility, accentedness, and fluency as we have operationalized these dimensions. After a few preliminaries, we discuss the steps in making good speech recordings and the preparation of audio materials for listening tasks. Rating and other judgment tasks are then covered, as is the effective administration of quasi-experimental listener tasks. A link to sample materials is provided, including PRAAT experiment files.

INTRODUCTION

This article is based on an invited workshop we offered at PSLLT 2019 at Northern Arizona University in Flagstaff. As seasoned researchers and regular reviewers for applied linguistics journals, we have accumulated experience and knowledge about the pitfalls of executing and publishing empirical research (Munro & Derwing, 2015). Drawing on the lessons we have learned, we designed the workshop to raise awareness among new researchers of potential problems and ways to avoid them. In addition to offering recommendations on procedures, we provided examples of tools for use in the specific types of research designs employed in many contemporary pronunciation studies. A link to some of the materials used in the workshop is provided here: www.sfu.ca/~mjmunro/courses/flagstaffworksho.html.

Initial Plan

One of the first steps in developing a research plan is to consider the type of data to be collected. A key question concerns the suitability of the data for the design of the study. That is, will the data address the research question or questions that you, the researcher, have defined? Choices to be considered include longitudinal and cross-sectional measures, data from before-and-after interventions, and qualitative, quantitative or mixed assessments.

Once a suitable design is selected, an additional question to consider is whether you can conceive of other studies for which further data could be collected at the same time. If you can think of ways to strategically utilize data collection time and other resources, you may be able to use your participants' time more efficiently, and even collect data for more than one study. A second question to ask is whether your prospective data could also be added to a corpus. As our field moves to 'big data' it is worthwhile to consult with existing corpora developers for guidelines. (Amanda Huensch and Shelley Staples, in particular, are currently planning on developing a corpus especially for pronunciation researchers.) As members of a research community, we can all benefit from maximizing our shared research resources. Keep in mind that you will need to cover these

possibilities in your Research Ethics Board (Institutional Review Board) proposal and your consent forms.

Preparation for Collecting Audio or Video Data

You may opt to do your audio or video recordings in your own research lab or at a remote site. The first choice offers the advantage of more controlled conditions; as a result, the acoustic quality of the materials is usually high. On the other hand, recording at a school or workplace can be much more convenient for research participants, and may be the only feasible way to conduct certain types of studies.

In either case, before making recordings of L2 speakers, be sure to plan extra time to build rapport with the participants and to explain the purpose and nature of the tasks. It is not always necessary to make a checklist (although if travelling off site, this is highly recommended), but at least verify that all materials are available and all equipment is functional.

For off-site data collection, it is best to have at least two team members to ensure a smooth outcome. In non-laboratory surroundings, it is likely that sooner or later something will go wrong (Bailey, 1983), such that one member will have to address the problem while the other remains with the experiment participants. In school settings, plan for extra time and delays in data collection due to such factors as unexpected classroom tests, field trips, bad weather, and illnesses. Consider all these possibilities for the site, so as to minimize wasted time.

When collecting data from students within a language program, clarify to the program coordinator and the affected teachers how your research may be useful in the future, even if it may not directly help the participants themselves. In studies involving multiple contact times, such as longitudinal investigations, be prepared to engage with your participants in personal interactions. This not only benefits *them*, but it can help *you* by enhancing participant retention. Our participants have, among other things, asked us to provide job references, assist them to address cultural and linguistic misunderstandings, and read over their university papers for grammar and spelling errors. Whenever possible, we complied.

When data are collected in the lab, many of the concerns of off-site recording are alleviated. Participants may need compensation for transportation costs (e.g., bus fare, parking), and just as in off-site recording, unexpected problems can occur (e.g., late arrivals that conflict with the next appointment, no-shows). Keep in mind that participants may need a map and detailed directions to locate your lab. Strategically placed signs in your building may be helpful. If you collect data on the weekends or in the evenings to accommodate the participants' schedules, be sure that the building entrances are not locked.

Eliciting Speech

Two issues that must be addressed with respect to elicitation are the target content (words, sentences, longer monologues, or interactions) and manner (reading, picture narratives, open-ended response, unrehearsed interaction) of your data. The first of these is often determined by the research questions; if you are studying prosody, you must collect connected speech, but if your

interest is vowels, your target items will be smaller units. Manner refers to the particular tasks used to collect speech data, as seen in Table 1. Notice the typical trade-off between convenience of data collection and *ecological validity*, the latter referring to applicability of research findings to the actual contexts of interest. For instance, while read-aloud materials are very convenient and allow the researcher to control the content completely, reading aloud in itself is not usually representative of the speech found in everyday interactions (Levis & Barriuso, 2012). Findings from such tasks may therefore be of little or no value in helping us understand a learner's pronunciation in typical day-to-day life.

Table 1
Types of Elicitation and their Validity and Convenience

Manner	Ecological Validity	Convenience
Reading aloud	Often poor	Good
Picture narrative	Can be good	Middling to challenging
Open-ended response	Depends; can be good	Middling to challenging
Unrehearsed interaction	Often very good	Challenging

Making Recordings

Ideally, recordings should be made in a truly quiet location, preferably a sound-treated room or booth. In an off-site room, the environment should be as quiet as possible, without such distractions as ringing phones, computer alerts or knocks on the door. (We suggest putting up a 'do not disturb' sign.)

One of the most important pieces of recording equipment is a good quality microphone connected to a computer or other recording device such as a stand-alone recorder, a tablet or a smartphone. Several free software packages are suitable for recording. Audacity™ (Audacity Team, 2019) is user-friendly and offers a wide range of valuable editing features. Praat (Boersma & Weenink, 2019), which is well-suited to many other functions in speech experiments, is not recommended for recording because of its limited interface. Apps on smartphones often give satisfactory audio recordings, though an external microphone is necessary, and only minimal editing features are usually available.

The accepted recording practice is to use monaural format, so that when the audio material is later presented to listeners through headphones, they hear exactly the same audio in each ear. Stereo recording adds nothing of value to most speech research; however, it doubles file sizes and may complicate later editing. For excellent audio quality use a sampling rate of 44100 Hz and a resolution of 16 bits. Before the experimental materials are recorded, an optimal recording level should be set using the software's "gain" adjustment. (The speaker should produce several sample items, thus becoming familiar with the speaking task.) Note that if the level is too loud, peak clipping will result in distortion, and if it is too soft, the recording will sound noisy when played back at a suitable volume. Either type of error can yield unusable audio recordings. Remember that every voice is different. It is poor practice to instruct participants to speak "louder" or "softer" than their natural volume because they will tend to fall back to their default patterns over the course of the recording task. Adjust the equipment, not the person!

Audio files should be recorded and saved in .wav format because of its universal accessibility. Most specialists advise against saving files in a compressed format such as .mp3 because of degradation of the audio. Note also that once a file has been compressed, it cannot be returned to its "undegraded" state, even with the best file conversion software.

After the recorded audio has been edited into separate files for analysis and judgments, the final step is to normalize the files to ensure a consistent volume from one audio file to the next. This is a mandatory step if the recordings are to be presented to listeners for ratings. Figure 1 illustrates a waveform of a speech recording in Audacity™ that has been peak normalized using the "Normalize" function in the "Effects" menu. Although other types of normalization are possible, we have found this approach to be the simplest and most effective in the vast majority of cases. In the options, select "Remove DC offset" to center the signal on the zero line. Normalize to the recommended level of -1 dB, which is just slightly below the maximum amplitude and ensures that no peak clipping is introduced.

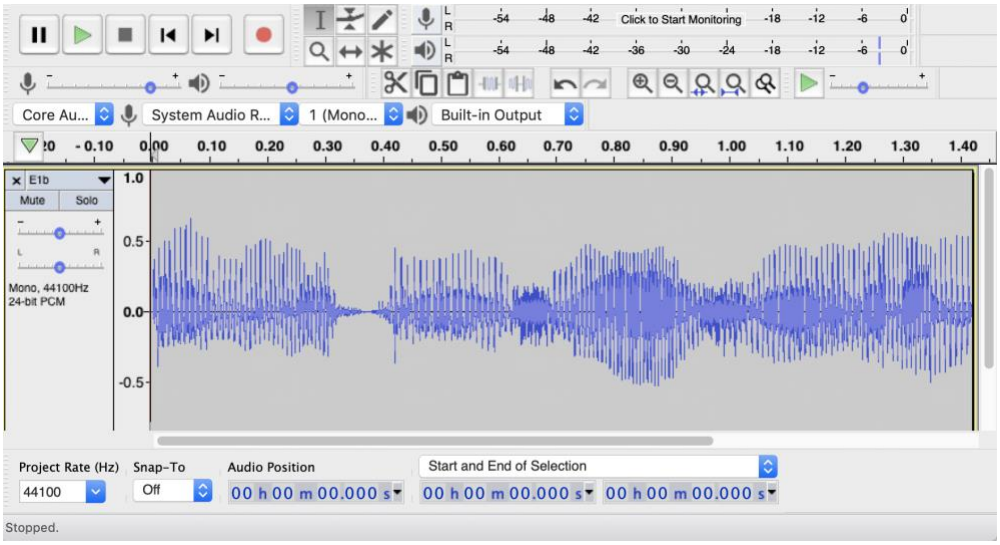

Figure 1. A normalized recording in the Audacity™ interface window.

Listening Tasks

Once a set of recordings has been prepared, an appropriate type of speech evaluation or analysis must be used. In much of our work, this step has entailed speech ratings, comparisons, or categorical judgments performed by either naïve listeners or linguistically trained judges (e.g., Munro & Derwing, 1995; Munro & Derwing, 2008). The listener-based approach is one of the most important and well-documented techniques in L2 speech research. Not only does it yield reliable, valid outcomes (Derwing & Munro, 2015), but it places the emphasis on human perceptions of speech. Rating data can therefore tell us something about how L2 speech is received and processed by a listening audience. Pronunciation researchers have a wide range of possible listening tasks at their disposal, including those listed in Table 2.

Table 2
Pronunciation Listening Tasks

Listening Task Type	Dependent Measure
Rating	Comprehensibility, Fluency, Accentedness, Irritation …
Orthographic Transcription	Intelligibility (words correct)
T/F Sentence Verification	Intelligibility; Response Time (processing difficulty)
Forced Choice Identification	Intelligibility of Vowels, Consonants, Words

Although we have also used acoustic analyses to complement listener-based data, we are highly skeptical of the use of acoustic data *alone*, despite claims for its "objectivity," as opposed to rating data, which some commentators characterize as "subjective." In fact, interpreting acoustic measures in terms of such critical dimensions as comprehensibility and intelligibility is very difficult, and as yet the relationship is poorly understood. A naïve assumption is that if a particular acoustic property of L2 speech is relatively close to that of native speech, then it can be interpreted as a sign of intelligibility. There is little reason to have faith in that assumption, however, because of the complex multiple-dimensional nature of speech acoustics. "Closeness" on one dimension may be nullified or overshadowed by "distance" on one or more other dimensions, including dimensions that the researcher has not measured or even considered. It is not surprising, then, that in one recent study (Chan & Hall, 2019), certain acoustic distances were found *not* to predict comprehensibility or intelligibility. In the final analysis, it is other human beings with whom L2 speakers interact, so human ratings provide the best window on the speech dimensions that interest us.

Ratings and Other Judgement Tasks

Traditionally, judgement tasks were done with pen and paper, usually in a group setting. This approach can sometimes still be justified, provided that a quiet location is available and that small groups of listeners hear different randomizations of the stimuli.

For individual collection of judgments, a variety of applications exist. One of the easiest to use is the free software Praat (Boersma & Weenink, 2019). Some sample experiment files for L2 pronunciation research are available at the following website: www.sfu.ca/~mjmunro/courses/flagstaffworksho.html. The PRAAT interface can be modified to collect ratings on any scale, e.g., comprehensibility, fluency, acceptability, irritability, or accentedness, with any desired number of response points. However, it is advisable to use a minimum of 9 points for global speech ratings (Munro, 2018) to ensure good reliability. Figures 2 and 3 show screen shots, first with Multiple Forced Choice (MFC) response buttons, and then with an MFC quasi-continuous scale. The latter type of scale is seen by the rater as a solid line with no numbered points even though it actually consists of 1024 separate buttons. Such a scale has been successfully used in comprehensibility studies (Crowther, Trofimovich, Saito & Isaacs, 2015).

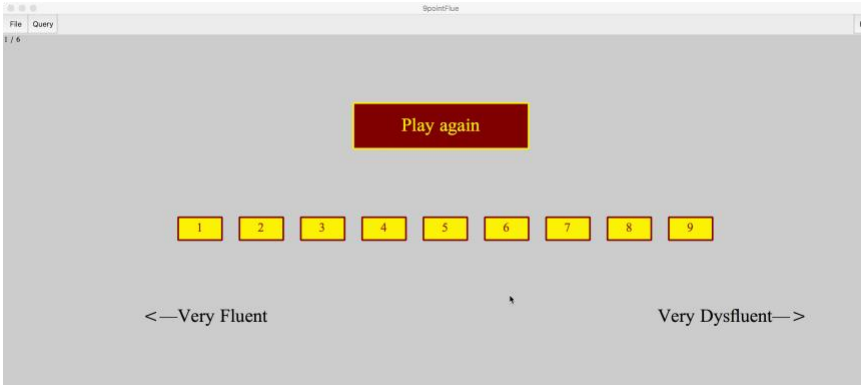

Figure 2. Sample Praat MFC screen with a discrete 9-point scale.

Figure 3. Sample Praat MFC screen with a quasi-continuous scale of 1024 points.

Another type of judgement task is categorical assessment. For instance, listeners may hear a series of words beginning with /p/ or /b/ and be asked to judge which consonant appeared at the beginning of each word. Another example is illustrated in Figure 4, which shows a screen used for assessing the local intelligibility of vowels. In this case the listeners heard individual words and selected the button corresponding to the vowel they perceived. Note that the listeners were linguistically sophisticated and could use IPA symbols, but key words in standard orthography could be used with naïve listeners. Similar experiments could also be designed for tone and stress categorizations.

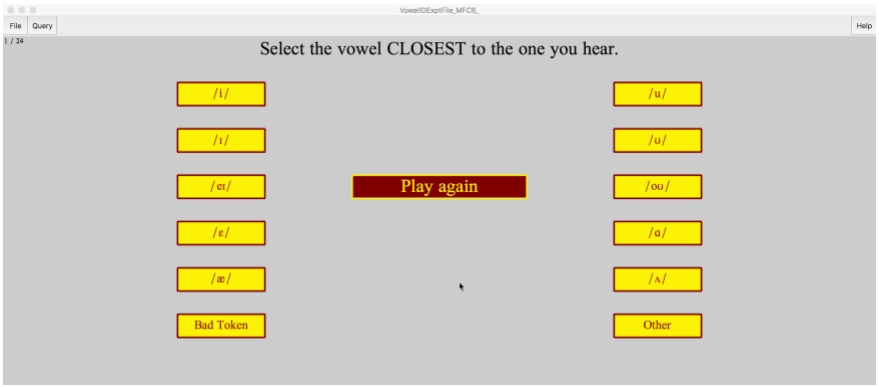

Figure 4. Sample Praat MFC screen for assessing local intelligibility of selected English vowels.

Researchers should carefully consider whether it is useful to include the optional replay button (and decide how many replays are allowed). An *oops* button can also be added, which repeats the last item to change the response, but this feature can sometimes confuse the listener.

Setting Up the Listening Task

Preparation of a listening task requires consideration of the variables shown in Table 3.

Table 3
Variables to Consider in Listening Task Preparation

Variable	Comments
individual stimulus length	Longer stimuli do not give "better" ratings. A *poor* strategy is to play a long stimulus (2 - 3 minutes) and then have the listener give a rating. In our work, we have found that 20 seconds is about the maximum amount of speech material that listeners can attend to. Many of our studies have used sentence-length utterances of only a few seconds each.
number of stimuli to be used/number of listening sessions per listener	The number of stimuli will dictate the number of sessions required of the listener. For large stimulus sets, it is advisable to break up stimuli into random subsets to be judged on different days.
pacing	Most rating tasks are best structured as self-paced, which means that the software presents a new stimulus item only after the listener has responded to the previous item. In some tasks (e.g., in response time studies) it is not acceptable for listeners to ponder their responses for indefinite intervals. In that

	case a maximum fixed inter-trial interval is appropriate. If the listener doesn't respond within the interval, no rating is collected, and the next stimulus is presented.
task length	Long tasks must include rest breaks. The total session should not exceed 1 hour. For an hour-long task, a three-minute break should be required, during which participants should be encouraged to drink water, eat something, and stretch.
warm-up items	A warm-up is necessary for task familiarization and volume adjustment. The warm-up stimuli should represent the range of items that will be judged. In some studies, the practice set consists of a random selection of the actual test items. However, if test items must be heard only once each during the entire task (as in a sentence verification task), a separate set of practice items is required.
randomization	Praat offers a number of randomization options. In most cases, each listener should hear the stimuli in a unique order that is different from that heard by all other listeners. If administering to small groups, ensure that each group hears a unique randomization.
catch trials	"Catch trials" are items that are not drawn from the actual stimulus set. These are advisable for pen and paper tasks to ensure that each listener has kept in step with the stimulus presentation. For instance, listeners may hear a stimulus that the researchers know to be highly comprehensible. After data collection if one rater assigns a stimulus a score of "7," while all other raters judge it a "1," the researcher may determine that the rater was not paying attention or fell out of step. Ratings on catch trials are not used in subsequent analyses.

Administration of Tasks

Any listening tasks, whether in small groups or individual format, should be carried out in the quietest possible conditions with a minimum of distractions. Listeners should be advised not to eat, chew gum, or drink anything while doing the task. Their phones should be turned off. Headphones should be used whenever possible; if not, high quality speakers are essential. A 'do

not disturb' sign should be placed on the door of the lab or the room where the task is conducted. Small groups should be advised to suppress giggling, coughs, and sneezes.

To ensure consistency across listening sessions, clear instructions should be presented from a script. These instructions should be neutral so as not to bias the raters (see Taylor Reid, Trofimovich & O'Brien, 2019). If a picture task was used to elicit productions from the L2 speakers, the listeners should view the pictures prior to completing the task. This reduces the effect of increasing familiarity with the content as the task progresses. Generally speaking, unsupervised take-home or online tasks are unsatisfactory because the researcher has no control over noise variables, distractions, or even the sobriety of the participants. In some crowd-sourcing techniques, such as Mechanical Turk (see Nagle, 2019), measures may be available to ensure that the participants comply with instructions. These techniques are recommended only if the researcher recruits substantially larger listener cohorts than in face-to-face data collection.

A Few Words on Data Analysis

We will not cover data analysis in this paper, but a couple of peripheral issues deserve mention. When carrying out statistical modelling, we suggest consulting recent papers in such journals as *Journal of Second Language Pronunciation*, *Language Learning*, *Studies in Second Language Acquisition*, and the *Journal of Phonetics*. Follow the contemporary practices for reporting inter-rater reliability and effect sizes.

Interpretation of Results

When interpreting the results of a typical rating study (e.g., of comprehensibility or some other dimension), it is a fallacy to assume that numerical ratings have an inherent, context-independent meaning. In fact, they are meaningful only in relation to each other because they are not criterion-referenced. The fact that a particular speaker receives a mean rating of "7" on a comprehensibility scale tells us nothing about that speaker in absolute terms. However, because well-designed comprehensibility studies normally show high reliability, it is nearly always safe to assume that a speaker who receives a mean rating of "7" differs in comprehensibility from someone who is rated as "3."

A common error in interpretation is a failure to look beyond group tendencies. To avoid over-generalizing, it is essential to closely examine individual speaker and listener performance. In our experience, some of the most interesting findings in L2 speech research have to do with individual differences or with patterns that emerge from examining subsets of the larger data corpus.

We also caution researchers not to overestimate the importance of "statistically significant" results when effect sizes are small. It is misleading to draw strong conclusions on the basis of weak effects. It is also unacceptable to assert that the findings of a study have implications for pedagogy unless such implications are clearly articulated by the researcher.

Miscellaneous Issues

It is sometimes necessary to incentivize participation in experiments. Your Research Ethics Board likely has guidelines for your institution's policies. Speakers may be willing to provide speech samples for a monetary reward or gift. Listeners from university classes may receive course credit or a monetary reward.

In pronunciation studies it is usually essential to administer a Language Background Questionnaire (LBQ) so that satisfactory personal profiles of participants are available. While no accepted standard exists, the slides posted on the accompanying website included suggestions on a number of matters that should be canvassed with participants. Often LBQ instruments are administered through a web interface (e.g., SurveyMonkey); check your own institution for privacy regulations.

ACKNOWLEDGMENTS

We acknowledge Ron Thomson's important contribution to many of the ideas presented here, and we thank Mary Grantham O'Brien and John Levis for numerous helpful comments on an earlier draft. Thanks also to the conference organizers and the editors of the proceedings.

ABOUT THE AUTHORS

Murray J. Munro is a Professor in the Department of Linguistics at Simon Fraser University, Vancouver, Canada, where he has taught linguistics and phonetics for over 25 years. His published books include *Pronunciation* (co-edited with J. Levis, 2017) and *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research* (co-authored with Tracey Derwing, 2015). His new book, entitled *Applying phonetics: Speech science in everyday life*, will be published by Wiley Blackwell in 2020. Murray's research centers on the ways in which linguistics can be used to address practical problems and has appeared in a wide range of journals covering the speech sciences and applied linguistics. mjmunro@sfu.ca

Tracey Derwing is a Professor Emeritus of TESL at the University of Alberta, and an Adjunct Professor in Linguistics at Simon Fraser University. With Murray Munro, she has extensively researched L2 pronunciation and fluency, especially the relationships among intelligibility, comprehensibility, and accent. In 2015, they co-published *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research.* Tracey has investigated native speakers' speech modifications for L2 speakers. She has also conducted workplace studies involving pragmatics and pronunciation. As a director of The Prairie Metropolis Centre of Excellence for Research on Immigration and Integration for eleven years, she has a deep interest in factors contributing to successful social integration of newcomers, most notably, the development of strong oral communication skills. tderwing@ualberta.ca

REFERENCES

Audacity Team (2019). Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 2.3.2 retrieved November 12th, 2019 from https://audacityteam.org/

Bailey, K. (1983). Illustrations of Murphy's Law abound in classroom research on language use. *TESOL Newsletter*, *17*, 4-5, 22-31.

Boersma, P. & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.1.06, retrieved November 12, 2019 from http://www.praat.org/

Chan, K. Y. & Hall, M. D. (2019). The importance of vowel formant frequencies and proximity in vowel space to the perception of foreign accent. *Journal of Phonetics*, *77*, 1-22.

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly, 49*, 814-837.

Derwing, T. M. & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam: John Benjamins.

Levis, J. & Barriuso, T. A. (2012). Nonnative speakers' pronunciation errors in spoken and read English. In. J. Levis & K. LeVelle (Eds.). *Proceedings of the 3rd Pronunciation in Second Language Learning and Teaching conference*. (pp. 187-194). Ames, IA: Iowa State University.

Munro, M. J. & Derwing, T. M. (1995). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning, 45*, 73-97.

Munro, M. J., & Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning, 58*, 479-502.

Munro, M. J. & Derwing, T. M. (2015). A prospectus for pronunciation research methods in the 21st century: A point of view. *Journal of Second Language Pronunciation, 1,* 11-42. DOI 10.1075/jslp.1.101mun.

Munro, M. J. (2018). Dimensions of pronunciation. In O. Kang, R. I. Thomson, & J. Murphy (Eds.) *The Routledge handbook of contemporary English pronunciation* (pp. 413-431). London & New York: Routledge.

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

Taylor Reid, K., Trofimovich, P., & O'Brien, M. (2019). Social attitudes and speech ratings: Effects of positive and negative bias on multiage listeners' judgments of second language speech. *Studies in Second Language Acquisition, 41*, 419-442. doi:10.1017/S0272263118000244