Received: 19 October 2019 | Revised: 26 February 2020 | Accepted: 10 April 2020

DOI: 10.1111/lnc3.12377

ARTICLE

WILEY

Using audio stimuli in acceptability judgment experiments

Yourdanis Sedarous | **Savithry Namboodiripad**

Department of Linguistics, University of Michigan, Ann Arbor, Michigan

Correspondence
Savithry Namboodiripad, Department of Linguistics, University of Michigan, Ann Arbor, MI.
Email: savithry@umich.edu

Abstract

In this paper, we argue that moving away from written stimuli in acceptability judgment experiments is necessary to address the systematic exclusion of particular empirical phenomena, languages/varieties, and speakers in psycholinguistics. We provide user-friendly guidelines for conducting acceptability experiments which use audio stimuli in three platforms: Praat, Qualtrics, and PennController for Ibex. In supplementary materials, we include data and R script from a sample experiment investigating English constituent order using written and audio stimuli. This paper aims not only to increase the types of languages, speakers, and phenomena which are included in experimental syntax, but also to help researchers who are interested in conducting experiments to overcome the initial learning curve. Video Abstract link: https://www.youtube.com/watch?v=GoWYY1O9ugs

1 | INTRODUCTION

This paper has two main audiences: For those who are already experienced with acceptability judgment experiments in particular, we provide motivation to use non-written stimuli. For those interested in conducting acceptability judgment experiments, particularly for under-described languages and varieties in non-lab contexts, we provide practical steps to get started.

Yourdanis Sedarous and Savithry Namboodiripad contributed equally, and should be considered joint first authors.

For all, we discuss best practices when conducting an experiment with audio stimuli, and we give examples using multiple experimental platforms.

Although acceptability judgments have gained popularity in recent years among syntacticians, the value of doing so has been (and often still is) debated (see Schütze & Sprouse, 2013 chp 3; Myers, 2009 for some overviews of these debates). The appropriate method(s) for investigating a particular phenomenon depends on a variety of factors; here, we assume that the decision to use an acceptability judgment experiment has already been made (as opposed to elicitation, corpus methods, other types of experimentation, etc.). We focus on spoken languages in our guidelines for implementation, but when it comes to motivating the use of non-written stimuli, the discussion is relevant to signed and spoken languages alike.[1]

The rest of Section 1 discusses contexts in which it might be beneficial or necessary to use non-written stimuli in acceptability judgment experiments, namely, increased representation of languages, varieties, speech communities, speakers, and empirical phenomena. We argue that "Widening the net"—to use the term from Anand, Chung, and Wagers (2011)—is not only possible, but crucial for our field to move forward. As part of this goal, we focus in particular on the practicalities of using audio stimuli in Section 2. Sections 3, 4, and 5 provide tutorials for setting up and distributing experiments using audio stimuli in Praat, Qualtrics, and PennController for Ibex, respectively. Section 6 refers readers to resources on data analysis, and we conclude in Section 7.

1.1 | Written stimuli stop us from moving beyond WEIRD participants and WISPy languages/varieties

Psychology experiments suffer from over-representation of WEIRD (Western, Educated, Industrialized, Rich, Democratic) participants, calling into question the validity of supposed cognitive universals (Henrich, Heine, & Norenzayan, 2010). Psycholinguistics, analogously, suffers from over-representation of WISPy (Written, Institutionally supported, Standardized, Prestige) languages and varieties, spoken mostly by WEIRD-M(onolingual) participants.[2] As Clancy and Davis (2019) discuss for biological anthropology, WEIRD often also means White; in the context of North America, for example, limiting participants and research questions to apply only to "Native English Speakers" can lead to a whiter participant pool. Further, for many varieties of English, studying phenomena specific to the written Standard means studying a particular White Standard (e.g., Bucholtz, 2001; Davila, 2012).

Anand et al. (2011) surveyed over 4,000 psycholinguistics conference and journal abstracts and found that "ten languages accounted for at least 85% of the research." Each of these languages has official status somewhere in the world, and only one of the 24 total listed is a historically minoritized language without official status (American Sign Language). The authors list suggestions for how to address this imbalance, including increased conversation and collaboration between fieldworkers and psycholinguists, support for researchers who speak understudied languages, and the development of more robust and flexible methods, tools, and approaches.

Here, we consider more widespread use of nonwritten stimuli as a way to address that final point. The majority of languages/varieties do not have a standardized writing system, and using audio stimuli allows for their inclusion in experimental linguistics. Even in communities which make use of written languages, written and spoken/signed varieties of languages can vary considerably in structure and associated social meaning. This includes contexts which are commonly labeled di/polyglossic, such as Egyptian Arabic–Standard Arabic, Haitian Creole–

French, spoken Tamil–written Tamil–English, American Sign Language–English, as well as contexts not usually labeled as such, like conversational and written Mainstream US English (Biber, 1993; Halliday, 1994).

This has additional relevance for high-contact varieties: While code-switching is a common multilingual practice, the degree to which it is also a written language phenomenon varies considerably across communities. Standardized writing systems associated with codeswitched varieties are rare, and many codeswitched combinations involve languages without converging alphabets. Written stimuli are just not possible in these cases. For example, Sedarous (2018) investigated the syntax of Egyptian Arabic-English codeswitched sentences. Her study focused on the acceptability of codeswitches within different locations in the construct state, a genitive phrase found in Semitic languages in which the possessor is directly adposed to, and functions as a single prosodic unit with, the possessed element. Using audio stimuli ensured that participants analyzed the construction as a single prosodic unit, and avoided potential infelicity associated with seeing two markedly different scripts.[3]

In addition to facilitating the inclusion of underrepresented languages/varieties in experimental linguistics, using audio stimuli allows for the inclusion of underrepresented speakers. Even in languages which have standardized writing systems, access to written language is not available to all. This could be due to age (i.e., children who have not yet learned to read), a general lack of access to education, or other pressures like displacement, globalization, (neo-)colonialism, or belonging to a minoritized community.

While research on the language use and comprehension of these speakers is important in its own right, and it can also lead to insights which would have otherwise gone unnoticed. For example, Namboodiripad, Kim, and Kim (2019) found gradient but not categorical differences in acceptability of constituent order when comparing three groups of participants: (a) *receptive bilinguals*, who grew up hearing but not speaking Korean in the United States, (b) *productive bilinguals*, who grew up hearing and speaking Korean in the United States, and (c) those who grew up speaking Korean in Korea. Written stimuli would have excluded the US groups, and a production study would have excluded the receptive bilinguals. Similarly, Scontras, Polinsky, Tsai, and Mai (2017) used audio stimuli to investigate scope ambiguity in English-dominant speakers of Mandarin. They found that these speakers lacked inverse scope in Mandarin (patterning with Mandarin-dominant speakers), and in English (diverging from self-reported "native" speakers of English). Again, this research would not have been possible with written stimuli.

1.2 | Audio stimuli and syntactic phenomena

Certain research questions require the use of non-written stimuli: For example, Ritchart, Goodall, and Garellek (2016) directly manipulated prosody in their experiment investigating the sources of the English that-trace effect. Šimík and Wierzba (2015) used audio stimuli to study the influence of stress on word order in Czech. Research on the effect of disfluencies or filled pauses on parsing (e.g., Lau & Ferreira, 2005) also necessitates the use of audio stimuli. This section covers advantages of using audio/signed stimuli more broadly, beyond cases for which the research question necessitates manipulation of audio.

The relationship between prosodic and syntactic phenomena has been an active area of investigation in spoken and signed languages. Investigations of processing and production of clausal boundaries (e.g., Wingfield, 1975 for English, Nicodemus, 2009 for American Sign

Language), left-periphery phenomena in Spanish (Sequeros-Valle, 2019), counter-sluicing in Japanese (Hiraiwa & Kobayashi, 2019), overt scope in Hungarian (Brody & Szabolcsi, 2003), and many more, discuss an important role for prosody. Discussions and analyses of constructed examples sometimes include notes on intonation, even if that is not a main part of the argument; indeed, many informal sources of data are from the spoken modality. As such, spoken or signed stimuli may be even more important for experimentalists who are interested in testing claims based on data from these types of sources.[4]

Written stimuli are not prosody-free; Breen (2014), in her review of literature relating to the Implicit Prosody Hypothesis (Fodor, 2002), discusses the considerable evidence that prosody and sentence processing are intertwined during reading. As Kitagawa and Fodor (2006) state that "[...] default prosody intrudes when no prosody is specified in the input. Thus judgments of visually presented sentences are not prosody-free judgments, but are judged as if spoken with default prosody." However, implicit prosody does not obviate non-written stimuli. Breen discusses studies which found individual differences in the positing of prosodic boundaries which in turn correspond to differences in high versus low relative clause attachment preferences (Jun, 2003; Swets, Desmet, Hambrick, & Ferreira, 2007). It is not possible to ensure that participants are positing the same "default" prosody, so controlling for prosody via audio/signed stimuli is preferable to the introduction of potential hidden variation. Of course, when constructing experimental items, prosody must be consistent within conditions and appropriate for the sentence (see Section 2 for tips).

Audio stimuli also allow for more direct comparisons of production and comprehension. For example, resumptive pronouns in English are relatively common, but they are consistently rated low in experimental contexts. Ferreira and Swets (2005) conducted an acceptability judgment experiment on resumptives using audio and written stimuli. They did not find a difference between the results in the two versions of the experiment, but they stated that, because these constructions "are not sentences that tend to occur in written English," it was necessary to use audio stimuli.

Depending on the context, written stimuli can bring forth prescriptive ideologies which researchers may like to avoid. To take one example, Beltrama and Xiang (2016) used audio stimuli in an acceptability experiment investigating resumptive pronouns as facilitators of island extracted sentences in Italian. They were interested in a particular regional variety of Italian, and they stated that audio stimuli would "increase the naturalness of the task." This relates to the finding that some speakers who may be less confident in a language tend to be more hesitant to reject constructions which other, more confident, speakers may accept (e.g., Orfitelli & Polinsky, 2017). Audio stimuli can help to make clear that it is everyday language which is under discussion and control for these hidden variables which are particular to written stimuli.

The use of audio stimuli rather than written stimuli in acceptability judgment experiments is relatively recent. While it is conceivable that the modality of stimuli presentation may have a main effect on the acceptability judgments (i.e., written stimuli could consistently be rated higher than audio stimuli, or vice versa), few studies have directly tested this.

1.3 | Interim summary

Our goal in this section was to outline some advantages of moving away from written stimuli. We do not intend to downplay the contributions that studies of written languages and processing in the written modality have made to (psycho)linguistics. However, as discussed by Henrich et al. (2010) and Majid and Levinson (2010), making generalizations about a particular

language or all languages from this type of data must be done with the proper caveats and contextualization.[5]

The arguments laid out here are in many cases well known among (psycho)linguists. Many of our colleagues do not need to be convinced that they should or must use audio stimuli; the barrier lies with *how* to implement audio stimuli in acceptability experiments. The remainder of this paper provides guidelines for carrying out these experiments, with the hope that this will lead to an increase in the types of language varieties, syntactic phenomena, and speech communities which are included in experimental syntax.

2 | GENERAL GUIDELINES FOR IMPLEMENTING AUDIO STIMULI IN ACCEPTABILITY EXPERIMENTS

This section provides general guidelines for recording stimuli, contextualizing the task, and choosing between platforms. We assume that contextually appropriate experimental stimuli and fillers with a range of acceptability have been created, assigned to lists, and (pseudo)randomized. For more on experimental design, see Sprouse and Schütze (2013) and/or Sprouse and Almeida (2017).

2.1 | Recording stimuli

Stimuli should be recorded in a soundproof or sound-attenuated booth, if available. If this is not possible, take care to ensure a quiet recording location. If there is some ambient background noise, ensure that it is uniform across conditions, as having background noise in one condition but not others could lead to an undesirable imbalance in how the items are rated.

Recording the stimuli using high-quality microphones is ideal. Some options include an AKG C 520 headset mic (approximately 220.00 USD), a Blue Yeti USB mic (approximately 110.00 USD), or a TONOR Q9 mic (approximately 50.00 USD). The first two microphones listed are used for phonetic research, which requires better sound quality than is necessary for the average acceptability judgment experiment; using a lower-quality microphone in a sound-attenuated booth can be adequate.

Files should be in `WAV` format for Praat experiments (described in Section 3) and `mp3` format if using the Soundcloud-Qualtrics integration (described in Section 4). Any format can be used with PennController for Ibex (described in Section 4.5). Recordings can be done in Audacity (audacityteam.org), a user-friendly software which is free and open-source, or directly in Praat. Audacity can save audio files in a number of formats, while Praat favors `WAV` format. There are a number of tutorials available online for recording in both of these programs.

Each item within a condition should have a consistent and natural intonational contour. If there is no research on prosody of the particular constructions you are investigating, you will need to rely on the intuition of the speaker who is recording the sentences. Here are some tips from our experience to help achieve consistency:

1. Record the sentences by condition (e.g., all SOV sentences together, all island-extracted sentences together, etc.)
2. Inhale and exhale between each sentence and start the next sentence anew (avoid a list intonation)

3. Say each sentence two or three times
4. Check the intonational contours of your sentences after the fact in Praat to ensure that they are consistent
5. Normalize loudness of files to control for volume mismatches (this can be done in Praat http://www.praatvocaltoolkit.com/normalize.html)[6]
6. Take care not to segment the sound files too close to the beginning or end of the sentence in order to avoid jarring onsets and offsets (make sure not to include the sound of the inhalation/exhalation)

Recording ungrammatical sentences can seem particularly tricky. Choose an intonational contour and keep it consistent for all sentences of the same type (e.g., all sentences which violate transitivity should have a similar prosody). The ungrammatical sentences should have natural-as-possible intonational contours; repeating the ungrammatical sentences a few times before recording can aid the recorder in creating consistent and relatively natural recordings.

The files can be segmented via Praat scripts (e.g.: `save_labeled_intervals_to_wav_sound_files.praat`). Using a transparent label which includes some combination of LIST, ID, ORDER (within the experiment), and CONDITION is helpful.

2.2 | Explaining the task

In order to contextualize the task for participants, it is useful to make clear the register of language you are interested in, such as everyday speech, the speech that they might use at home, the speech they might use at work, etc. As discussed in Section 1.1, this is especially relevant for participants who may have complex relationships with the languages, varieties, and/or sentence-types used in the experiment. When going into the experiment, participants likely will not be thinking about completely ungrammatical sentences in the range of sentences they expect to hear, and giving examples beforehand of the range of sentences participants should expect—from ungrammatical word-salad to perfectly acceptable unremarkable sentences—can put into context what is meant by acceptability.

If relevant, it is especially helpful to have a sentence with a construction which is known to be disprefferred prescriptively and explicitly state that, while these types of sentences are said to be "not proper" (or whatever terminology is most appropriate for the context), they are perfectly fine in everyday conversation and for the purposes of the experiment. In order to avoid the influence of explicit instruction on the experimental items, make sure any example sentences are different in structure from the experimental items. We have found these types of explanations especially helpful when working on non-Standard varieties or with language practices such as codeswitching which may be otherwise stigmatized, though it is a good practice overall.

Depending on the context, references to marking/grading as done in formal schooling might be helpful to introduce the idea of a rating scale, though it is important to state that participants should not mark the sentence for "correct" grammar, as may have been done on exams. In general, having as much information as possible about what ideologies participants may have about experimental items, whether through schooling or other means, is crucial for experimental design, framing how participants should approach the task, and (potentially) interpreting results. If the investigators are not members of the community themselves, or if they do not have access to this information, working with community members in the design or pilot phase will be especially necessary.

Here is an example for acceptability judgment experiments in Malayalam, conducted in Kerala, India in 2014 and 2016 by the second author. Participants were literate in English and Malayalam (with considerable variation in self-reported comfort with each), and there is awareness in the community of how the written Standard variety differs significantly from spoken language: *ezhuttubhaasha* "written language," *saahityabhaasha* "poetic language," and *sadaabhaasha* "everyday-language" are commonly used terms. The scope of the study was described by mentioning each variety, discussing how each is legitimate and of interest for linguistic inquiry, and stating that the present study was about everyday language. Participants were asked to consider each sentence on its own, think about how it would sound if they heard someone say it, and rate the sentence accordingly. In addition, participants were given examples of a "normal" sentence (such as "I like ginger") and a nonsense sentence (a jumble of words such as "Father ball yesterday"). While the first few sentences of the experiment also contained fillers which had a range of acceptability to reinforce the point, stating examples out loud helped to demonstrate what was meant by "unacceptable."

Beyond questions necessary for participant screening, asking questions about language experience, attitudes, and background at the end of the experiment is preferable in order to avoid potentially priming participants. If it is logistically possible, excluding participants after the fact can help with this.

2.3 | Deciding between platforms

Where are your participants located? Are you looking to run your experiment online? What is your level of comfort with coding? These are just some of the considerations at play when deciding on the platform to run your experiment. Before going into the details of implementing the experiment in each platform, this section addresses some of the considerations. We consider Praat, Qualtrics, and PennController for Ibex in turn.[7]

Praat is free, familiar to many linguists, and the particular script discussed here, ExperimentMFC, has documentation. While Praat is a scripting language, the script for this experiment requires minimal coding, and the output of the experiment is a tab-delimited file that is easy to work with. Important for our goal of "Widening the net," Praat experiments can be run on a laptop and do not require an internet connection. However, as far as we know, these experiments cannot be run online, so the experiment must be conducted in person. In addition, the script included in the supporting materials of this paper does not randomize stimuli.

Qualtrics is subscription-based software which is primarily used for creating surveys, but is adaptable for creating acceptability judgment experiments. Qualtrics allows for easy online distribution, it integrates well with sites such as Prolific.ac, and it runs smoothly on mobile devices. However, for those without institutional access, subscriptions can be expensive. Finally, while distributing the experiment and collecting data is a quick process, setting up the actual experiment itself and manually cleaning the output is time-intensive.

PennController for Ibex (Zehr & Schwarz, 2018) is a JavaScript library designed and maintained by Jeremy Zehr and Florian Schwarz. It serves as an extension of Ibex Farm (Drummond, 2016), designed by Alex Drummond, making it possible to script experiments that utilize a wide range of features that were unavailable in Ibex (crucially for us, the use of multimedia files as stimuli). PennController for Ibex is user-friendly and highly customizable. Like Qualtrics, experiments created via PennController for Ibex are easily distributed online. Unlike Qualtrics, PennController for Ibex is free to use and an open-source tool that does not require a subscription fee.

3 | ACCEPTABILITY JUDGMENT EXPERIMENTS IN PRAAT

This section outlines how to conduct an acceptability judgment experiment using ExperimentMFC in Praat (Boersma & Weenink, 2019). Documentation can be found at http://www.fon.hum.uva.nl/praat/manual/ExperimentMFC.html; and the exact file used for the experiment discussed in this section can be found at https://github.com/ccc-lab/PraatAcceptability. The script is a text file which can be edited in any text-editing program.

As a reminder, the items should already be in pseudorandom order at this point. Enter the filenames (no need to include the `WAV` part) in pseudorandom order inside of the quotes, as seen in Figure 1. Do not delete the empty double quotes after each file name. Make sure to enter the correct "number of different stimuli" in the ExperimentMFC script (Figure 2). As necessary, change the introduction text to include the language that you are using, as seen in Figure 3. Praat supports Unicode, so you can include instructions in non-Roman alphabets.

Create a folder for each list (label *List 1*, *List 2*, etc.). Within each of those folders, create a folder called *sounds*, and put the corresponding sound files in each list. Put the edited script in the *List* folder; you should have a folder labeled *List X* with the script (labeled *List X*) and a folder labeled sounds which contains all of the `WAV` files for each list.

To run the experiment, load the script as a Praat object, select it, and hit *Run*. You must manually export the data before running that experiment again; Praat will override the data from the previous experiment once you hit *Run* again. So, if you have six lists and are running participants evenly across the lists, export the data before cycling back. The output is a tab-delimited file (`txt`) which has the following columns: *SUBJECT*, *STIMULUS*, *RESPONSE*, *REACTION TIME*, as seen in Figure 4.

SUBJECT is actually the version of the experiment (here, List 3), *STIMULUS* is the label for the sound file which was entered between the apostrophes, *RESPONSE* is the 1–7 rating given by the participant, and *REACTION TIME* is the time from the end of the stimulus until the

```
    "PracticeGood1" ""
"PracticeBad1" ""
"PracticeGood2" ""
"PracticeBad2" ""
"287" ""
"VSO15" ""
"340" ""
"SVOAgree23" ""
"315" ""
```

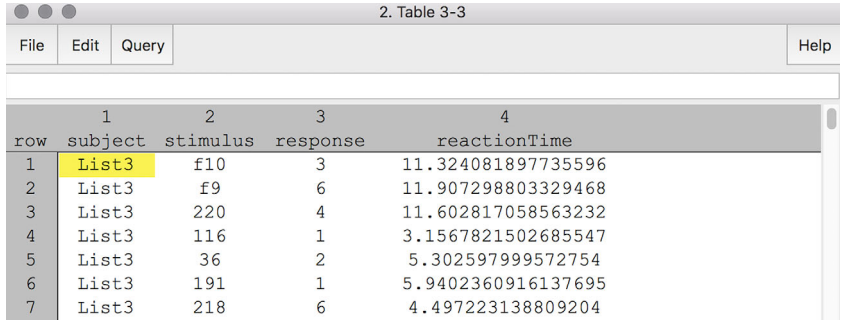
FIGURE 1 Filenames in the ExperimentMFC script

```
numberOfDifferentStimuli = 70
```

FIGURE 2 Where to edit the number of different stimuli in the ExperimentMFC script

```
startText = "You will be presented with some possible sentences of Malayalam.
For each sentence, rate it on a scale from 1 (very bad) to 7 (very good)
based on how it sounds to you as a native speaker of the language.

Some things to keep in mind:

-Give your first reaction. Don't try to analyze the sentence.

-There are no correct answers. We want to know how the sentence sounds to YOU.

Just be sure to rate each sentence on its own,
regardless of how simple or complicated it seems.

Ready? Click to begin."
```

FIGURE 3 Sample introduction text in the ExperimentMFC script



	1	2	3	4
row	subject	stimulus	response	reactionTime
1	List3	f10	3	11.324081897735596
2	List3	f9	6	11.907298803329468
3	List3	220	4	11.602817058563232
4	List3	116	1	3.1567821502685547
5	List3	36	2	5.302597999572754
6	List3	191	1	5.9402360916137695
7	List3	218	6	4.497223138809204

FIGURE 4 Sample output of Praat experiment

rating was entered.[8] With this setup, the researcher must manually keep track of the participant ID associated with each set of results.

The script included in the supplementary materials does not allow participants to hear sentences multiple times; this setting can be modified by adding a "replay button"; instructions can be found on the ExperimentMFC documentation online. For single-listen experiments, the rating screen comes up after the sound file plays, and participants may give their ratings via the number keys or by clicking the radio buttons with the cursor.

4 | ACCEPTABILITY JUDGMENT EXPERIMENTS IN QUALTRICS

In order to create a new project within Qualtrics, select the *New Project* tab under *Projects*. Assuming you do not have a template,[9] make sure to select *Blank Survey Project* to create a new project, and provide it with a relevant name.

4.1 | Multiple-listen acceptability

This section outlines how to create an experiment which allows participants to listen to each sentence multiple times before judging its acceptability.

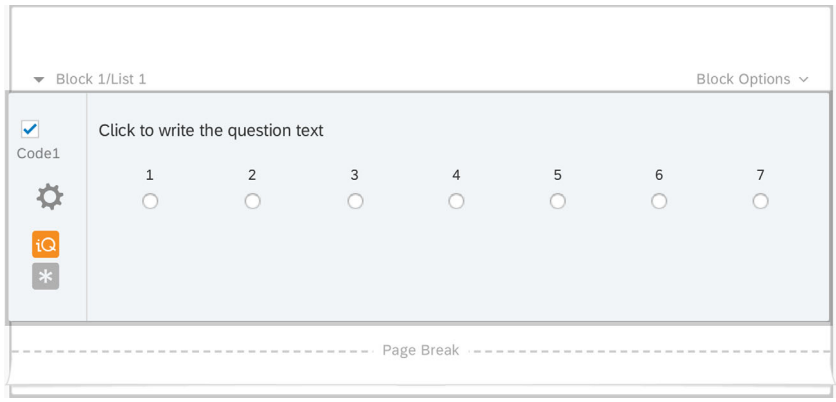


FIGURE 5 Sample question in Qualtrics survey

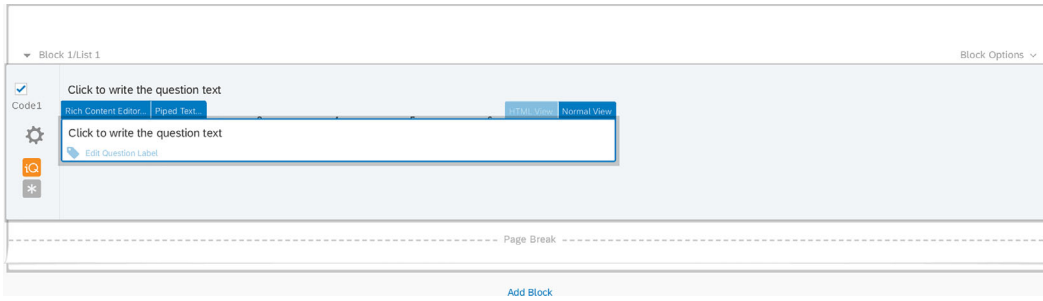


FIGURE 6 Rich content editor within Qualtrics survey

Upload sound files, which must be in `mp3` format, into your Qualtrics Survey *Library*. Create one "question" for each sentence. This question will include both the sound clip and the 1–7 rating scale. Name the question with a relevant stimulus tag for exporting purposes (e.g., Code1); participants will not see this tag when taking the experiment. This can be seen in Figure 5. Finally, set the *Question Type* to *Multiple Choice* with the appropriate amount of *Choices* selected.

Embed the sound clip by selecting the *Rich Content Editor* option within the question (Figure 6). Select the video icon option to insert media (Figure 7), and then choose the appropriate sound file from your library through *Select File From Library* (Figure 8).

Set *Answers* to *Single Answers* in order to limit participants to one response per item, and select the *Force Response* option in *Validation Options* so that participants cannot move on to the next item without providing a judgment.[10] Finally, select the *Add Page Break* option under *Actions* to present items one by one. An example of the full settings discussed in this section can be seen in Figure 9.

4.2 | One-listen acceptability

This section outlines how to create an experiment which is more analogous to the Praat experiment, allowing participants to hear items only once.

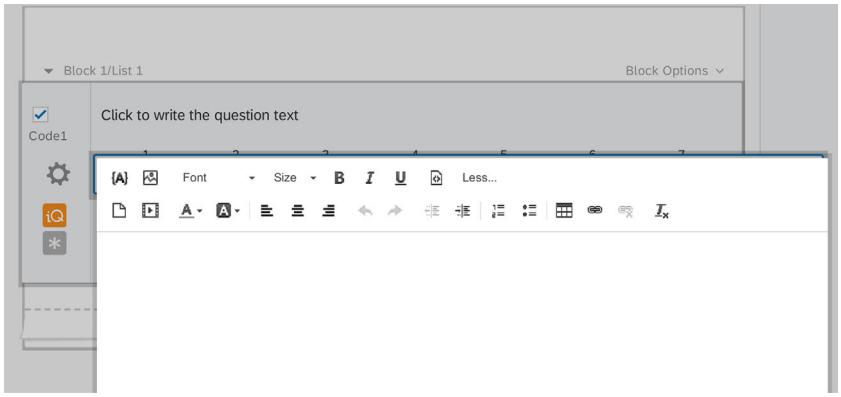


FIGURE 7 Selecting multimedia presentations in Qualtrics survey

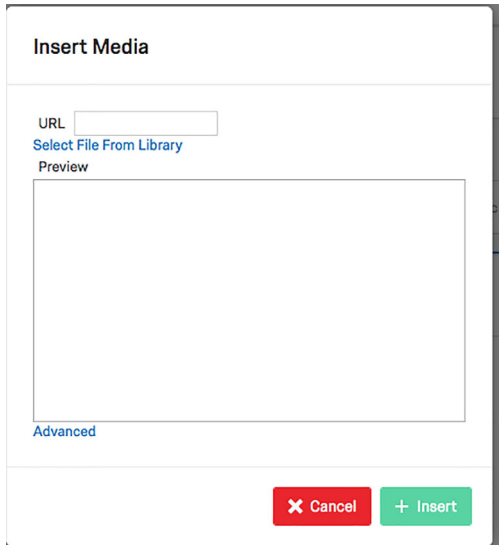


FIGURE 8 Inserting multimedia presentations in Qualtrics survey

Since Qualtrics does not provide an option to autoplay, sound files must be uploaded onto a hosting server such as Soundcloud.[11] For each item, you'll need to create three "questions" in the Qualtrics survey. The first question serves as the timer, the second question contains the sound clip, and the third question will serve as the 1–7 rating scale.

For the first question, select *Timing* from the *Change Question Type* option, and set the *Auto-advance after (seconds)* to the appropriate time depending on the length of your sound file, as seen in Figure 10. For example, if your sound file is 2 s long you may choose to set your timer to auto-advance after 3 s.

Embed each audio file by getting its .html link from Soundcloud. Check to make sure that the *Enable automatic play* option is selected in Soundcloud (Figure 11), select *HTML View* from the "edit" option of the second question in your Quatlrics survery, and paste the html code for the audio file . Alter the height dimensions so that your participants do not

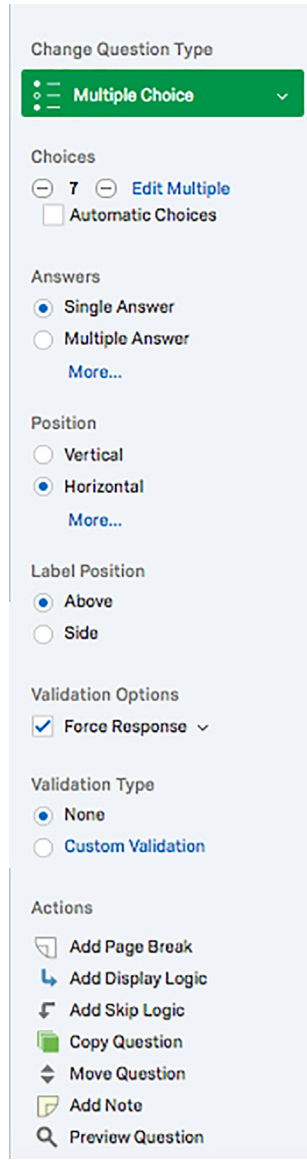


FIGURE 9 Sample settings in Qualtrics survey

see the item's file name. An example can be seen in Figure 12; there, the height was set to 33 pixels.

After setting the timing for the first question, and embedding the sound clip for the second question, add a page break by selecting the *Add Page Break* option under *Actions*.

For the third and final question associated with your item, insert the 1–7 rating scale as described in Section 4.1. Finally, add a page break after the third question.

In the end, your set of three questions should resemble Figure 13. Repeat this process for each item.

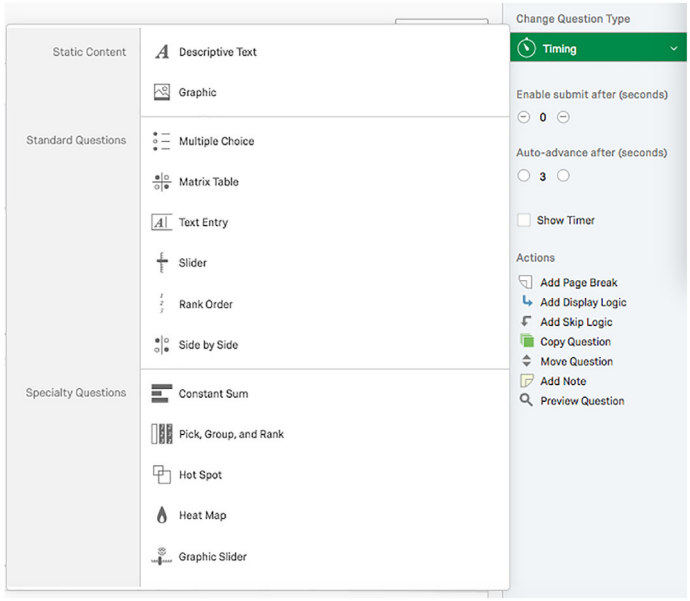


FIGURE 10 First question in Qualtrics: Timing

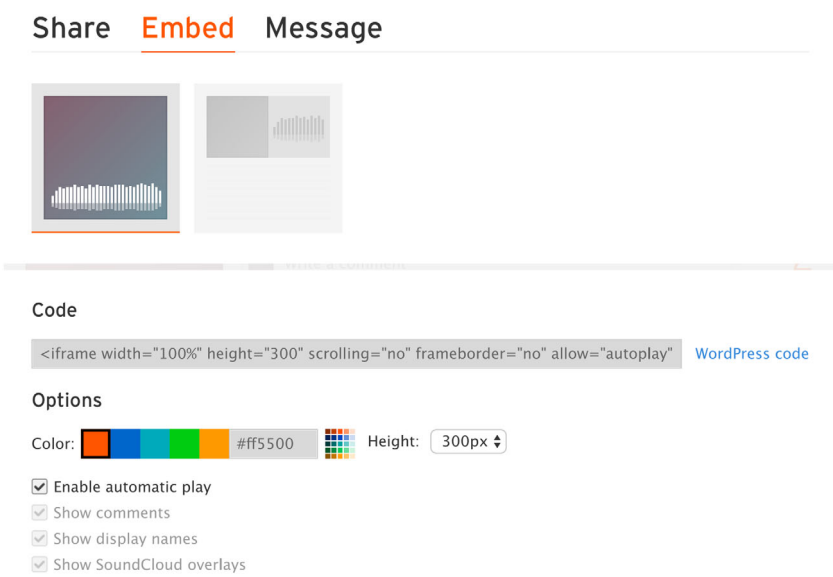


FIGURE 11 Getting the HTML link from SoundCloud

4.3 | Adding and randomizing blocks

For experiments with multiple lists, each list should be housed in a separate block. Create multiple blocks by clicking the *Add Block option* on the bottom of the survey. To randomize the questions in each block, click the *Block Options* and select *Question Randomization*. From there you can randomize the questions as necessary.

Randomizing blocks allows you to evenly distribute lists across participants. To do this, select the *Select Survey Flow* option under the *Survey* tab, which will show you a schematization

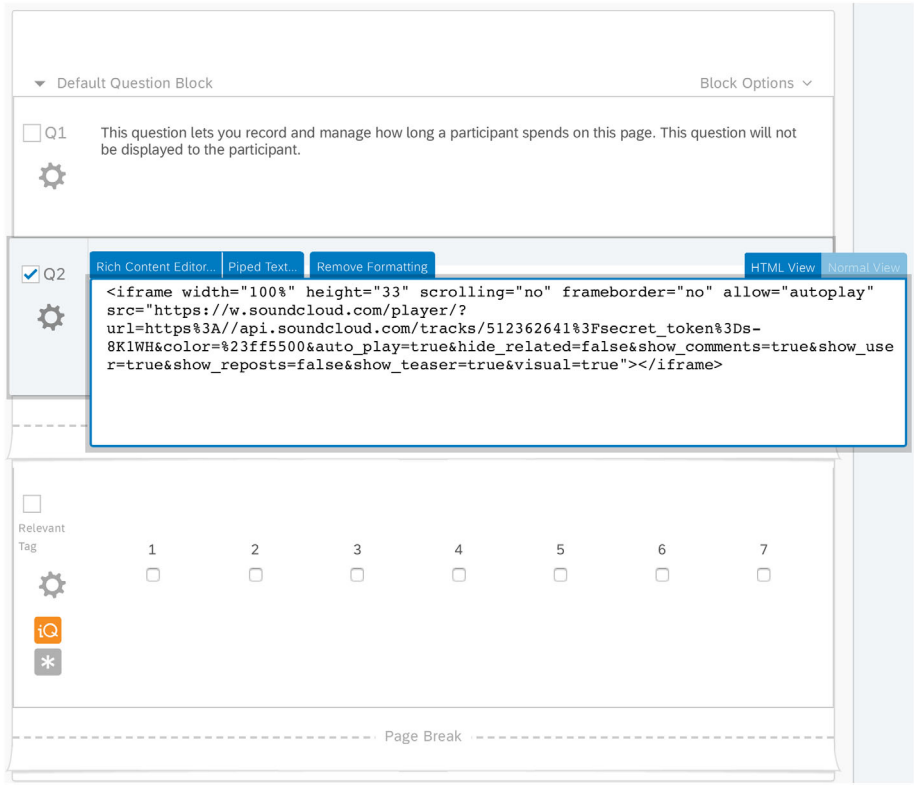


FIGURE 12 Embedding the HTML SoundCloud link into Qualtrics

of your experiment as a whole. Then in *Select Survey Flow* click *Add a New Element Here*, then select *Randomizer*. Move the different blocks (i.e., lists) into the *Randomizer*. To make sure that each participant hears only one list, set *Randomly present* ____ *of the following elements* option to *1*. To make sure that each list is seen by an equal number of participants, mark the *Evenly Present Elements* option.

4.4 | Distributing and data output

The Qualtrics experiment you created is now mobile and can be taken on any computer or phone. To send it out, select the *Distributions* option and choose the appropriate form of distribution that is relevant to your experiment.

After data collection is complete, export your results by going to the *Data & Analysis* tab in your Qualtrics experiment. Export your results (in `csv` format) from the *Data* subtab.

5 | ACCEPTABILITY JUDGMENT EXPERIMENTS USING PENNCONTROLLER FOR IBEX

This section provides a tutorial for creating a one-listen audio acceptability judgment experiment with a 1–7 Likert scale using PennController for Ibex. This platform is very customizable;

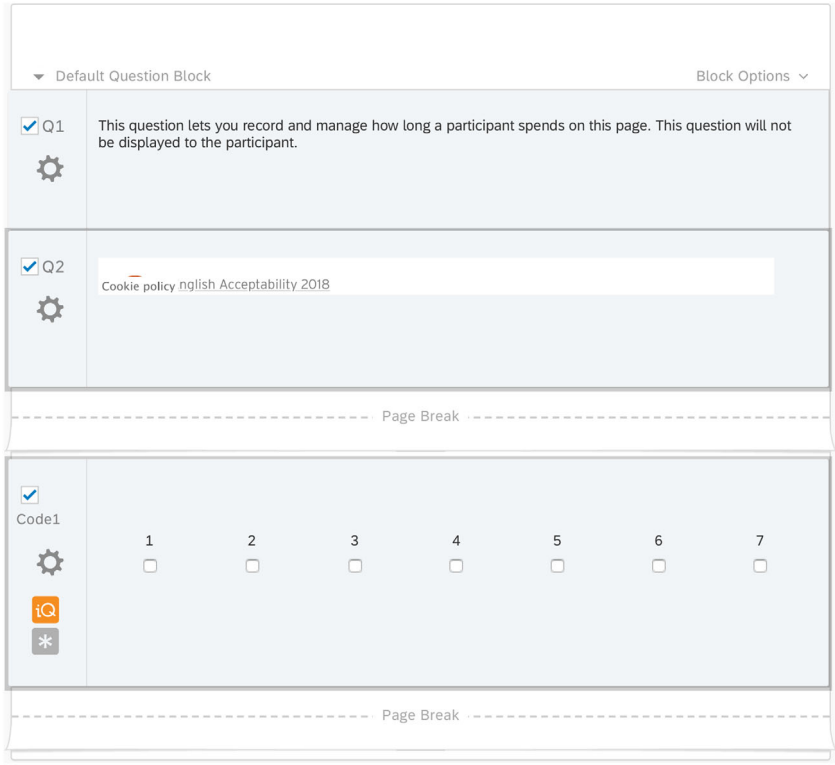


FIGURE 13 The three total questions for one-listen acceptability judgment experiments within a Qualtrics survey

for more options, information, and tutorials, visit the Penn Controller for Ibex Farm website: https://www.pcibex.net/documentation/.

If you do not already have one, first create a user account at https://expt.pcibex.net/login, then create a new experiment with a relevant name for your project. After creating a new experiment, you should be redirected to your experiment homepage, as seen in Figure 14.

Preload the audio files that will be used for your experiment as well as your pseudorandomized[12] `csv` file into Resources. An example of labeling the columns for your `csv` file can be seen in Figure 15, which includes example stimuli from Experiments 1 and 2 (see Supplementary Materials Data S1 and S2). PennController for Ibex understands that the items under *Group* indicate different lists, and so the program will only display one list to each participant.

Preload all relevant materials into *Resources* before continuing with the script. As of now, PennController for Ibex requires that you upload each sound file individually which can be time consuming for experiments using a large number of files. We suggest loading a zip file with all of the sounds at once. Follow this tutorial to batch-preload a zip file into *Resources*: https://www.pcibex.net/wiki/penncontroller-preloadzip/.

After preloading your files into *Resources*, upload the JavaScript file found in the supporting materials into *Script*. Edit the *welcome* and *demographic* texts as relevant. The data is output as a `csv` file.

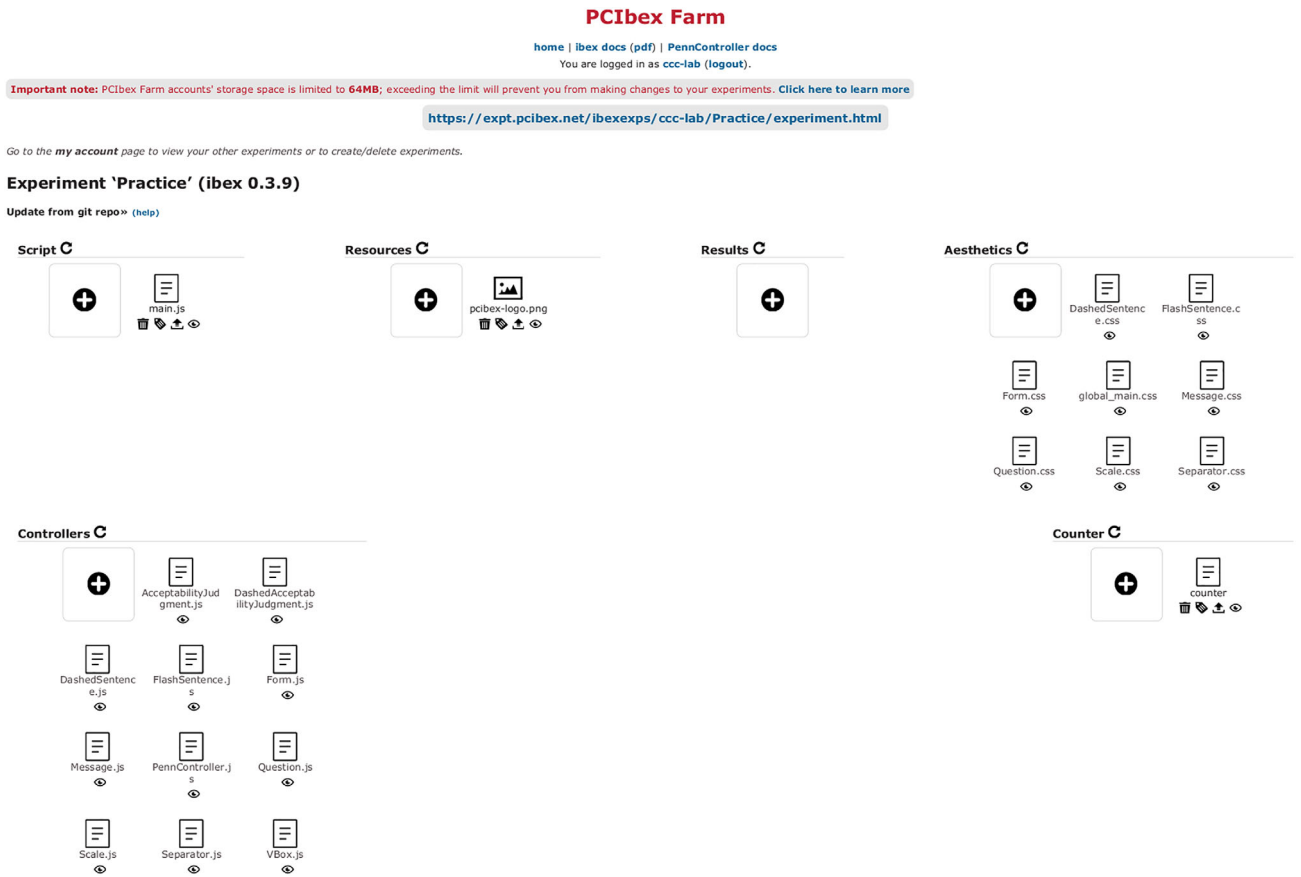


FIGURE 14 Experiment homepage for PennController for Ibex

Group	Exp Order	ID	Condition	Audio File	Label	Description
1	1	F18	FILLER	F18	FILLER	What do you think was on the table yesterday?
1	2	F9	FILLER	F9	FILLER	Jeff follow souring artichoke.
1	3	4	SOV	SOV_S11	SOV	a caddy a pencil sharpened
1	4	240	ShortBare	ShortBare_Sam	ShortBare	Who thought Sam bought the shoes?
1	5	3	SOV	SOV_S12	SOV	a butler a door opened
1	6	F26	FILLER	F26	FILLER	The keyboard that he touched a cord when looking at.
1	7	261	LongDOf	LongDOf_Frank	LongDOf	Which of the singers did you suspect Frank cooked for?
2	1	F18	FILLER	F18	FILLER	What do you think was on the table yesterday?
2	2	F9	FILLER	F9	FILLER	Jeff follow souring artichoke.
2	3	211	ShortBare	ShortBare_Jane	ShortBare	Who knew Jane demolished the buildings?
2	4	F11	FILLER	F11	FILLER	What would the girl could the tiger suddenly do?
2	5	34	SVO	SVO_S6	SVO	a waitress peeled a banana
2	6	294	ShortDOf	ShortDOf_Tom	ShortDOf	Which of the coachs believed Tom faked the injury?
2	7	88	VSO	VSO_S26	VSO	towed a trucker a trailer
3	1	3	F18	F18	FILLER	What do you think was on the table yesterday?
3	2	3	F9	F9	FILLER	Jeff follow souring artichoke.
3	3	3	220	ShortBare_Steve	ShortBare	Who discovered Steve produced the movie?
3	4	3	116	VOS_S6	VOS	Peeled a banana a waitress.
3	5	3	F30	F30	FILLER	When did the notebook that it sold brownies when writing on?
3	6	3	36	SVO_S11	SVO	A caddy sharpened a pencil.
3	7	3	F14	F14	FILLER	What will Tom buys when he could went on vacation?

FIGURE 15 Sample `csv` file organization to be uploaded under Resources in PennController for Ibex experiment

6 | PLOTTING AND DATA ANALYSIS

As the present focus is on issues particular to audio stimuli, for space considerations, we do not focus here on issues of data analysis: There are many available resources, and they apply equally to written and audio stimuli. Here, we point the reader to a few which may be especially helpful. Section 8 of Gibson, Piantadosi, and Fedorenko (2011) discusses analysis in R, and refers to other resources for information on R and statistical analyses. We highlight one crucial best practice from their tutorial here, which is the importance of visualizing the data using a histogram before conducting any analyses. An additional resource is Sprouse and Almeida (2017), which provides a more detailed discussion of hypothesis testing and statistical analysis (in the context of investigating false negatives, or type II errors, in acceptability judgment experiments).

The supplementary materials for this paper (n.b., using different scripts and packages from those Gibson et al. (2011) discuss) include sample data and an R script which can be used to walk through the process of data visualization. This script is intended to be user-friendly and accessible to those with even limited experience using R.

7 | CONCLUSION

Our aim in this paper was to provide arguments for the use of non-written stimuli in acceptability judgment experiments, as well as to provide researchers with the tools to implement these experiments. In doing so, we highlighted the benefits of using audio stimuli, namely expanding the representation of languages, speakers, and phenomena that are currently underdescribed in the psycholinguistics literature. We also supplied readers with step-by-step instructions for conducting acceptability experiments using audio stimuli via three platforms: Praat, Qualtrics, and PennController for Ibex. We hope that this contribution will encourage and aid our colleagues in using audio stimuli in their work.

ACKNOWLEDGMENTS

The authors would like to thank Carolin Carruth, Alexander Forrest, Elaine Francis, Hayley Heaton, Ryan Lepic, Amanda Richart-Scott, Vineet Tummala, and students in the 2019 LSA Summer Institute class: *Acceptability judgments in syntax: Theoretical interpretation and experimentation outside the lab*.

ORCID

Yourdanis Sedarous https://orcid.org/0000-0002-1968-1227
Savithry Namboodiripad https://orcid.org/0000-0002-7685-5895

ENDNOTES

[1]See Emmorey (1993), MacSweeney et al. (2002), Morford and Carlson (2011), Hosemann, Herrmann, Steinbach, Bornkessel-Schlesewsky, and Schlesewsky (2013), et alia for examples of acceptability judgment experiments with video stimuli in sign languages.

[2]See also Dahl's presentation (2015) on LOL languages (Literate, Official, and Lots of Users), and MYALS (Monolingual, Young, Available, and Literate Speakers), as discussed by Polinsky (Polinsky).

[3]Koronkiewicz and Ebert (2018) investigated the effects of stimulus modality on the judgments of Spanish-English codeswitched sentences and reported no significant effects for presentation modality. This is not surprising; Spanish and English have more congruent scripts, and these communities differ in their conventions around codeswitching in the written modality.

[4]As a reviewer pointed out, this could have contributed to non-replications of empirical claims, such as the non-replication of the "third wh-phrase effect" (Kayne, 1983) by Fedorenko & Gibson (2010).

[5]Bender and Friedman (2018) outline analogous issues as relevant for the field of Natural Language Processing. They propose including data statements as a best practice, and they argue that this will improve science by providing important context for results and, importantly, helping to mitigate and uncover bias and exclusion.

[6]Check the sound files afterwards to make sure that the sentence still sounds natural.

[7]Erlewine and Kotek (2016) discuss *turktools*, which integrates with Amazon Mechanical Turk. We do not discuss it here because, while audio stimuli are possible on this platform, the implementation requires more knowledge of coding than we are assuming, and the integration of audio stimuli is not documented. Also, we are hesitant to use Amazon Mechanical Turk as a platform for distributing experiments because of documented labor violations (Hara et al., 2018); for online distribution, we prefer sites such as Prolific.ac (Palan & Schitter, 2018), which performs additional participant screening and requires researchers to pay minimum wage (as determined in the United Kingdom). See Peer, Samat, Brandimarte, and Acquisti (2017) for direct comparisons of data quality on Amazon Mechanical Turk, CrowdFlower, and Prolific.ac; they found higher data quality, less participant dishonesty, and more participant diversity on Prolific.ac as compared to the other sites.

[8]N.b., In noisy settings and for unspeeded tasks, reaction time is not interpretable.

[9]Unfortunately, unlike with PCIBEX and Praat, we cannot provide a template for Qualtrics in the supplementary material. Since Qualtrics is a subscription-based survey maker there is no easy export function. Please contact the authors via email for access to an existing experiment in this format.

[10]Some aesthetic options: Setting the *Position* option to *Horizontal* yields a Likert scale which ascends from left to right. For numbers to appear above the radio button, set *Label Position* to *Above*.

[11]Make sure to upload the sound files as *private* rather than *public*.

[12]Unless PennController for Ibex is prompted to randomize items within a list, it will display them to participants in the order they are placed in the `csv` file. There are several methods for randomizing your experimental and filler items, which can be found in Section 6 of the Ibex 0.3.8 Manual: http://spellout.net/latest_ibex_manual.pdf.

REFERENCES

Anand, P., Chung, S., & Wagers, M. (2011). *Widening the net: Challenges for gathering linguistic data in the digital age*. Response to NSF SBE, 2020.

Beltrama, A., & Xiang, M. (2016). Unacceptable but comprehensible: The facilitation effect of resumptive pronouns. *Glossa*, *1*(1), 1.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, *19*(2), 219–241.

Boersma, Paul & Weenink, David (2019). Praat: Doing phonetics by computer [Computer program]. Version 6.1.05. Retrieved from http://www.praat.org/

Breen, M. (2014). Empirical investigations of the role of implicit prosody in sentence processing. *Language and Linguistics Compass*, *8*(2), 37–50.

Brody, M., & Szabolcsi, A. (2003). Overt scope in Hungarian. *Syntax*, *6*, 19–51. https://doi.org/10.1111/1467-9612.00055

Bucholtz, M. (2001). The whiteness of nerds: Superstandard English and racial markedness. *Journal of Linguistic Anthropology*, *11*(1), 84–100.

Clancy, K. B., & Davis, J. L. (2019). Soylent is people, and WEIRD is white: Biological anthropology, whiteness, and the limits of the WEIRD. *Annual Review of Anthropology*, *48*, 169–186.

Dahl, Ö. (2015). *How WEIRD are WALS languages*. Leipzig: MPI-EVA.

Davila, B. (2012). Indexicality and "standard" edited American English: Examining the link between conceptions of standardness and perceived authorial identity. *Written Communication*, *29*(2), 180–207.

Drummond, A. (2016). Ibexfarm. https://github.com/addrummond/ibexfarm

Emmorey, K. (1993). Processing a dynamic visual—Spatial language: Psycholinguistic studies of American sign language. *Journal of Psycholinguistic Research*, *22*(2), 153–187.

Erlewine, M. Y., & Kotek, H. (2016). A streamlined approach to online linguistic surveys. *Natural Language & Linguistic Theory*, *34*(2), 481–495.

Fedorenko, E., & Gibson, E. (2010). Adding a Third Wh-phrase Does Not Increase the Acceptability of Object-initial Multiple-wh-questions. *Syntax*, *13*(3), 183–195. http://dx.doi.org/10.1111/j.1467-9612.2010.00138.x

Ferreira, F., & Swets, B. (2005). The production and comprehension of resumptive pronouns in relative clause "Island" contexts. In A. Cutler, W. Klein, & S. Levinson (Eds.) *Twenty-First Century Psycholinguistics: Four Cornerstones* (263–278). Erlbaum.

Fodor, J. D. (2002). *Psycholinguistics cannot escape prosody*. Speech Prosody 2002, International Conference.

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*, *5*, 509–524.

Halliday, M. A. K. (1994). Spoken and written modes of meaning. In *Media Texts: Authors and Readers* (Vol. *7*, pp. 51–73).

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018, April). *A data-driven analysis of workers' earnings on amazon mechanical turk*. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, ACM. p. 449.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29.

Hiraiwa, K., & Kobayashi, Y. (2019). Countersluicing. *Syntax*. http://dx.doi.org/10.1111/synt.12190.

Hosemann, J., Herrmann, A., Steinbach, M., Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2013). Lexical prediction via forward models: N400 evidence from German sign language. *Neuropsychologia*, *51*(11), 2224–2237.

Jun, S. A. (2003). Prosodic phrasing and attachment preferences. *Journal of Psycholinguistic Research*, *32*(2), 219–249.

Kayne, R. (1983). Connectedness. *Linguistic Inquiry*, *14*, 223–249.

Kitagawa, Y., & Fodor, J. D. (2006). Prosodic influence on syntactic judgments. In G. Fanselow, C. Fery, R. Vogel, & M. Schlesewsky (Eds.), *Gradience in grammar: Generative Perspectives*. (336–358). Oxford University Press.

Koronkiewicz, B., & Ebert, S. (2018). Modality in code-switching research: Aural versus written stimuli. In L. López (Ed.), *Code switching: Experimental answers for theoretical questions* (pp. 147–193). New York, NY: John Benjamins.

Lau, E. F., & Ferreira, F. (2005). Lingering effects of disfluent material on comprehension of garden path sentences. *Language and Cognitive Processes*, *20*(5), 633–666.

MacSweeney, M., Woll, B., Campbell, R., McGuire, P. K., David, A. S., Williams, S. C., ... Brammer, M. J. (2002). Neural systems underlying British sign language and audio-visual English processing in native users. *Brain*, *125*(7), 1583–1593.

Majid, A., & Levinson, S. C. (2010). WEIRD languages have misled us, too. *Behavioral and Brain Sciences*, *33* (2–3), 103–103.

Morford, J. P., & Carlson, M. L. (2011). Sign perception and recognition in non-native signers of ASL. *Language Learning and Development*, *7*(2), 149–168.

Myers, J. (2009). Syntactic judgment experiments. *Language and Linguistics Compass*, *3*, 406–423.

Namboodiripad, S., Kim, D., & Kim, G. (2019). English dominant Korean speakers show reduced flexibility in constituent order. *Proceedings of CLS*, *53*.

Nicodemus, B. (2009). *Prosodic markers and utterance boundaries in American sign language interpretation*. Washington D.C.: Gallaudet University Press.

Orfitelli, R., & Polinsky, M. (2017). When performance masquerades as comprehension. In M. Kopotev, O. Lyashevskaya, & A. Mustajoki (Eds.), *Quantitative Approaches to the Russian Language*, (1–18). London: Routledge.

Palan, S., & Schitter, C. (2018). Prolific. Ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Peer, E., Samat, S., Brandimarte, L., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.

Polinsky, M. (in press). Experimental syntax and linguistic fieldwork. In J. Sprouse (Ed.), *Oxford Handbook of Experimental Syntax*. London: Oxford University Press.

Ritchart, A., Goodall, G., & Garellek, M. (2016). Prosody and the That-Trace Effect: An Experimental Study. In *Proceedings of the 33rd West Coast Conference on Formal Linguistics* (pp. 320-328).

Schütze, C., & Sprouse, J. (2013). Judgment data. In R. Podesva & D. Sharma (Eds.), *Research methods in linguistics, Chp. 3* (pp. 27–50). Cambridge: Cambridge University Press.

Scontras, G., Polinsky, M., Tsai, C. Y. E., & Mai, K. (2017). Cross-linguistic scope ambiguity: When two systems meet. *Glossa: A Journal of General Linguistics*, *2*(1), 1–28.

Sedarous, Y. (2018). *Code-Switching in the Egyptian Arabic Construct State*: *A Syntactic and Psycholinguistic Analysis*. Ms., University of Michigan, Ann Arbor.

Sequeros-Valle, J. (2019). The intonation of the left periphery: A matter of pragmatics or syntax? *Syntax*, *22*, 274–302. https://doi.org/10.1111/synt.12182

Šimík, R., & Wierzba, M. (2015). The role of givenness, presupposition, and prosody in Czech word order: An experimental study. *Semantics and Pragmatics*, *8*, 3–1.

Sprouse, J., & Almeida, D. (2017). Design sensitivity and statistical power in acceptability judgment experiments. *Glossa*, *2*(1), 14.

Swets, B., Desmet, T., Hambrick, D. Z., & Ferreira, F. (2007). The role of working memory in syntactic ambiguity resolution: A psychometric approach. *Journal of Experimental Psychology: General*, *136*(1), 64–81.

Wingfield, A. (1975). The intonation-syntax interaction: Prosodic features in perceptual processing of sentences. In *Structure and process in speech perception* (pp. 146–160). Chicago: Springer, Berlin, Heidelberg.

Zehr, J., & Schwarz, F. (2018). PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832

AUTHOR BIOGRAPHIES

Yourdanis Sedarous is a fourth year PhD candidate in the Linguistics Department at the University of Michigan, Ann Arbor. She is interested in syntax, multilingualism, and language contact, with a special focus on the Afro-Asiatic language family. She is currently exploring the processing effects of various syntactic structures within different populations of Egyptian Arabic-English bilinguals. Before starting her PhD work, she earned her BA and MA from The Ohio State University.

Savithry Namboodiripad is an Assistant Professor in the Linguistics Department at the University of Michigan, Ann Arbor, where she heads the Contact, Cognition, and Change Lab. Her research explores how language contact shapes language use, variation, and emergence. She uses experimental methods to explore syntactic typology, particularly in the domain of constituent order, and takes traditional experimental methods outside of the lab. She also studies various aspects of language contact in Malayalam, including language attitudes and maintenance practices in immigration and globalization contexts. She holds a BA and MA from the University of Chicago, and a PhD from the University of California, San Diego.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Sedarous Y, Namboodiripad S. Using audio stimuli in acceptability judgment experiments. *Lang Linguist Compass*. 2020;14:e12377. https://doi.org/10.1111/lnc3.12377