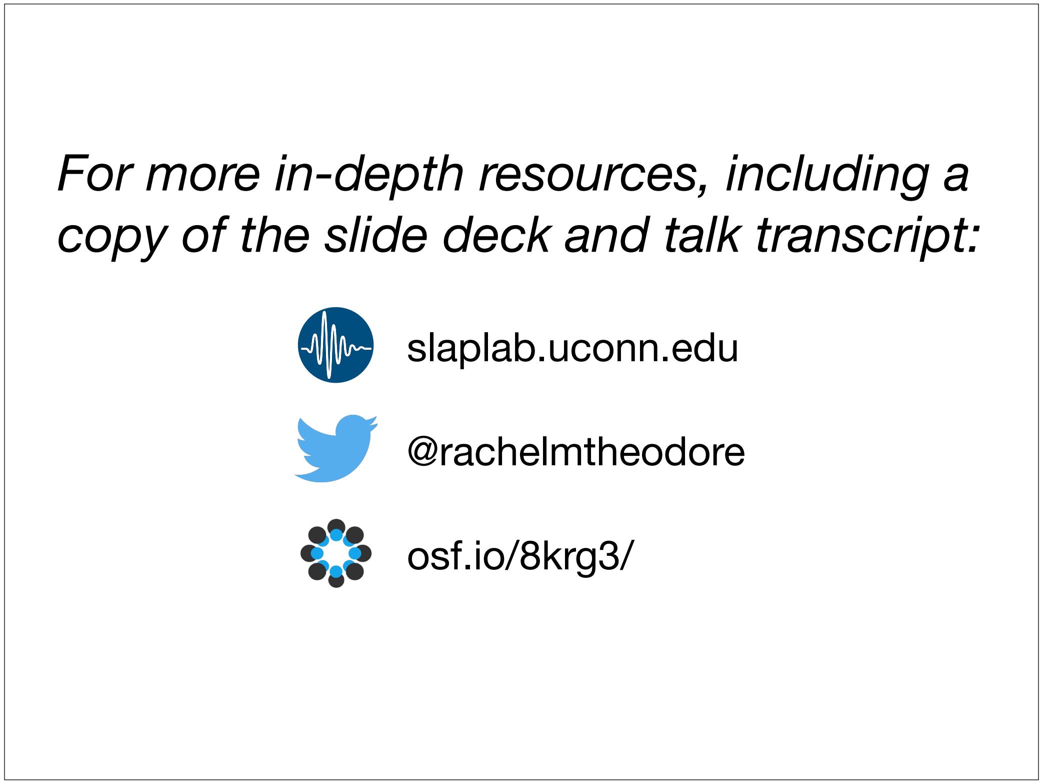
Critical considerations for conducting web-based speech perception research

Rachel M. Theodore, Ph.D.
Department of Speech, Language, and Hearing Sciences
Connecticut Institute for the Brain and Cognitive Sciences
University of Connecticut

Hello, and welcome to the thematic session, “Changes in Space: Online Experimentation.” I am excited to share with you:

Critical considerations for conducting web-based speech perception research.



For more in-depth resources, including a copy of the slide deck and talk transcript:

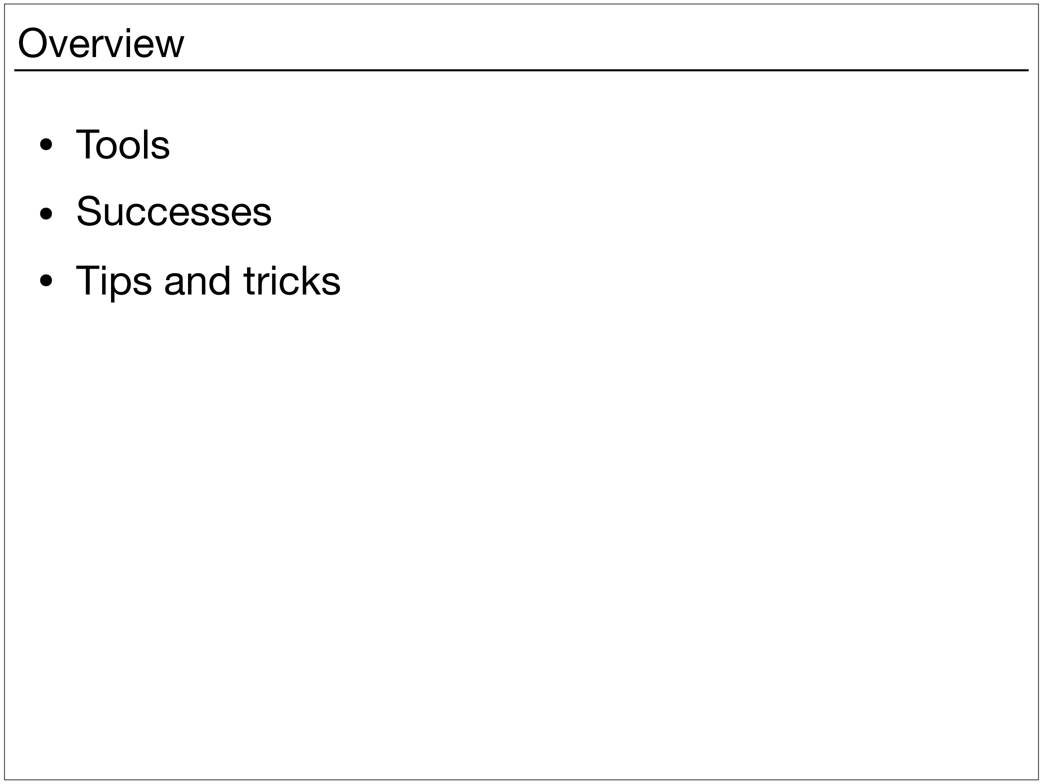
slaplab.uconn.edu

@rachelmtheodore

osf.io/8krg3/

So much so, that I've packed quite a bit into this 25-minute talk. I'd like to point you to some offline resources that you can find at our lab website, including a copy of this slide deck and a transcript of the talk.

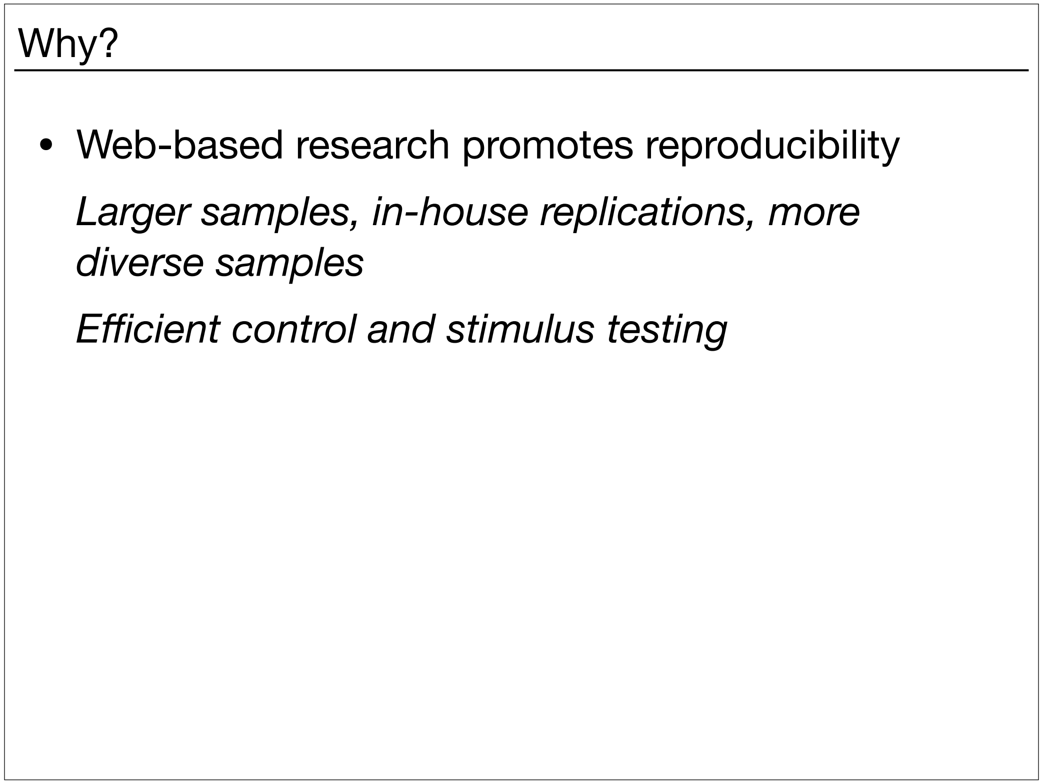
At our OSF page, you can find data and analysis code for all of the examples that I'll share with you today, and many other web-based experiments.



Overview

- Tools
- Successes
- Tips and tricks

In this talk, I’m going to share some tools, successes, and tips and tricks for promoting high quality web-based speech perception research. But first, I’d like to briefly talk about the why - as in, why would we ever consider abandoning the carefully controlled laboratory setting with our fancy headphones and sound booths and participants that we can actually see?



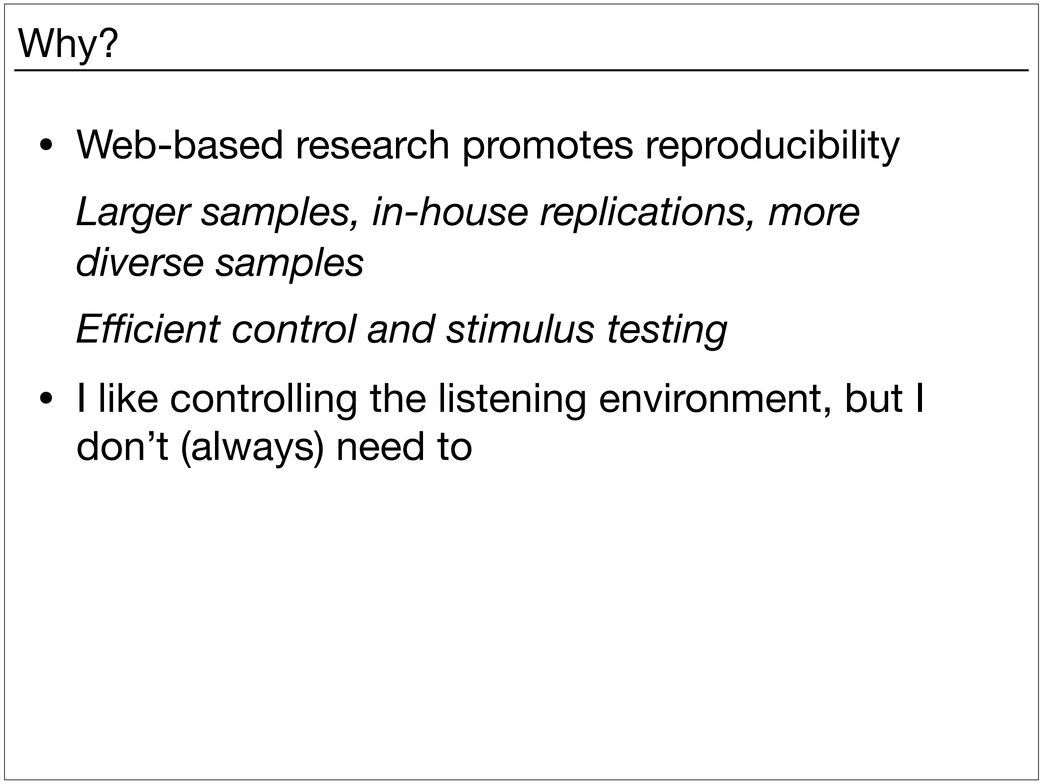
Why?

- Web-based research promotes reproducibility

  Larger samples, in-house replications, more diverse samples

  Efficient control and stimulus testing

For me, there are myriad reasons to do so. My foray into web-based research was in direct consequence of me completely drinking the reproducibility Kool-Aid. As we moved to adopt emerging best practices for promoting reproducibility of research, we needed to find ways to collect data from larger sample sizes. And make in-house replication studies the norm. And not limit our samples to reflect the demographics of our university. And run more control experiments. And better vet our stimulus sets. And verify that our results aren't contingent on a single stimulus set. And so on.



Why?

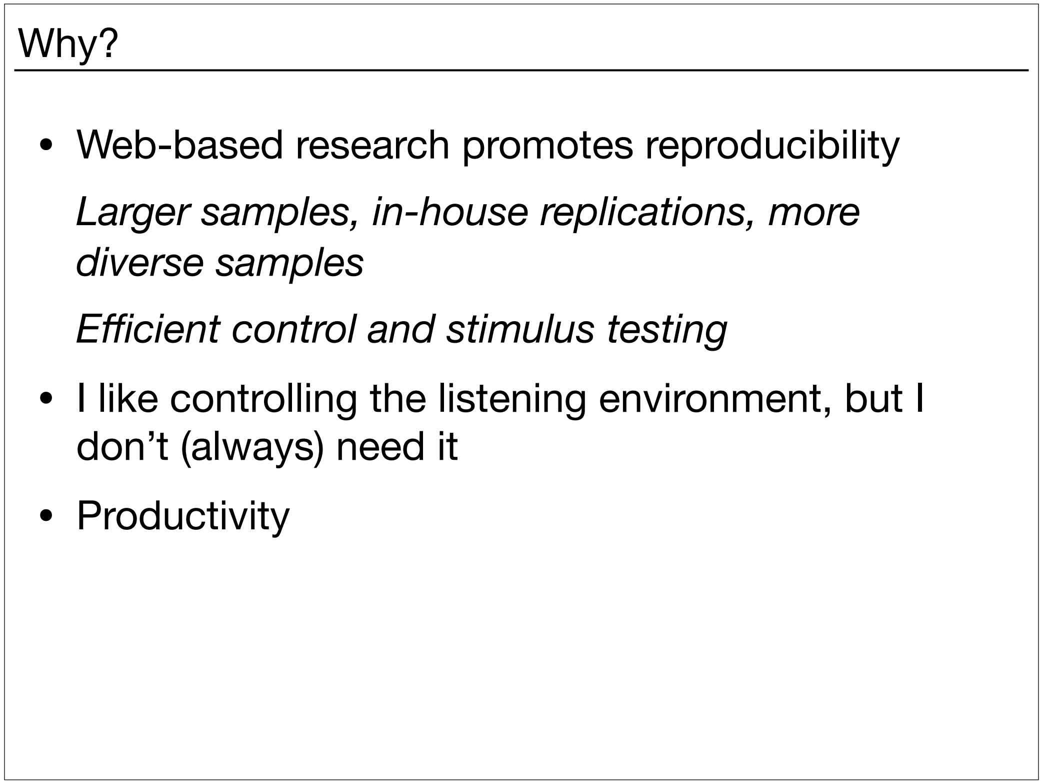
- Web-based research promotes reproducibility

  Larger samples, in-house replications, more diverse samples

  Efficient control and stimulus testing

- I like controlling the listening environment, but I don’t (always) need to

Many of our studies don't actually require a laboratory level of control over the listening and response environment. We don't present auditory stimuli at threshold levels. Processing time effects that we're interested in exceed keyboard timing error. I like to think that the things we study (and thus the things we claim) might actually be relevant in a more natural listening environment...



Why?

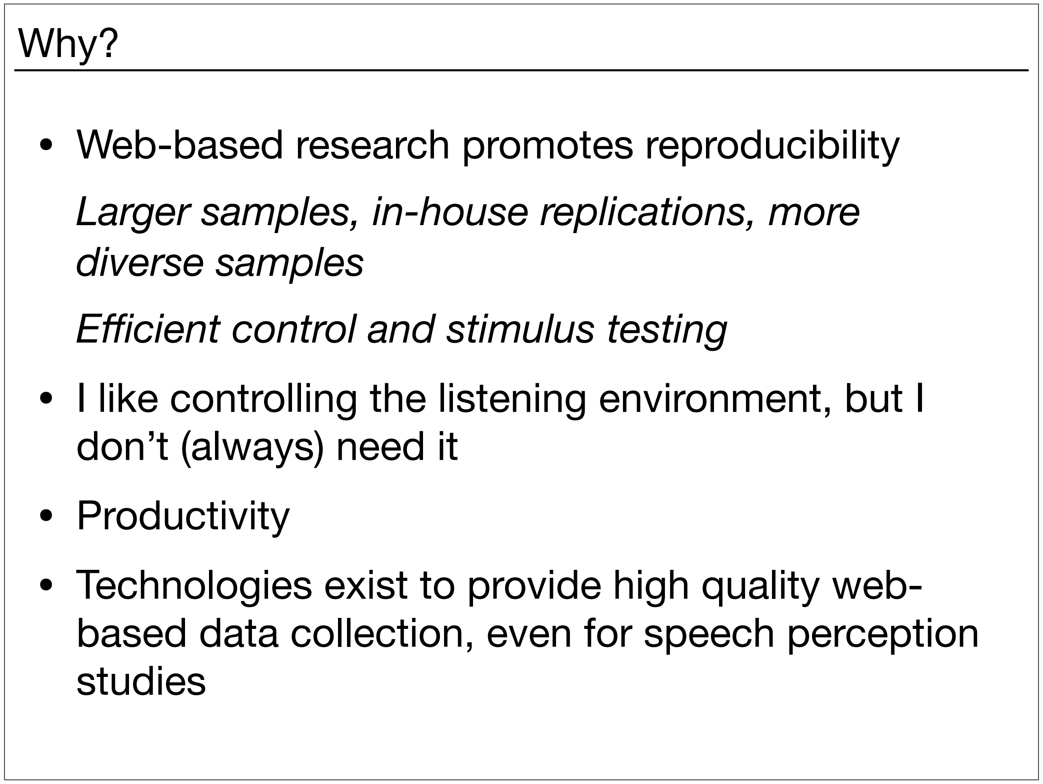
- Web-based research promotes reproducibility

  Larger samples, in-house replications, more diverse samples

  Efficient control and stimulus testing

- I like controlling the listening environment, but I don’t (always) need it
- Productivity

It's hard to keep up with my extremely productive colleagues if we limit testing to our physical lab. For better or worse, science moves very quickly these days. Online data collection also facilitates productivity of our trainees, who are increasingly expected to have strong publication records.



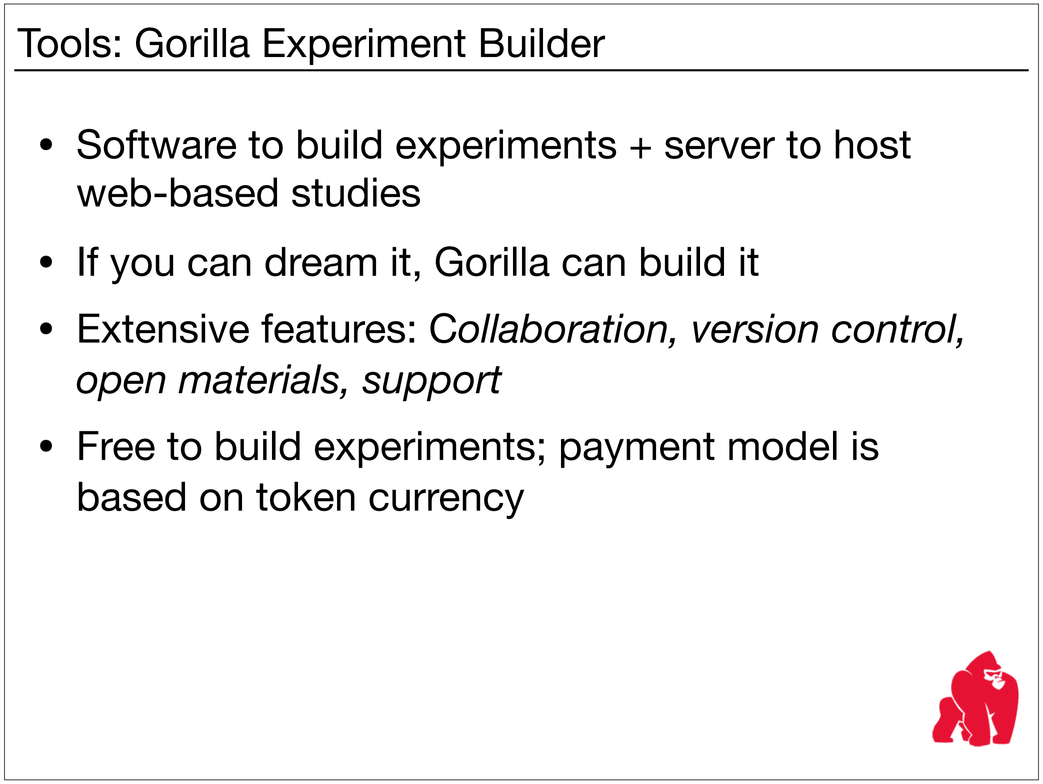
Why?

- Web-based research promotes reproducibility

  Larger samples, in-house replications, more diverse samples

  Efficient control and stimulus testing
- I like controlling the listening environment, but I don't (always) need it
- Productivity
- Technologies exist to provide high quality web-based data collection, even for speech perception studies

And finally, emerging technologies do exist to provide high quality web-based data collection, even for speech perception experiments.



Tools: Gorilla Experiment Builder

- Software to build experiments + server to host web-based studies
- If you can dream it, Gorilla can build it
- Extensive features: *Collaboration, version control, open materials, support*
- Free to build experiments; payment model is based on token currency

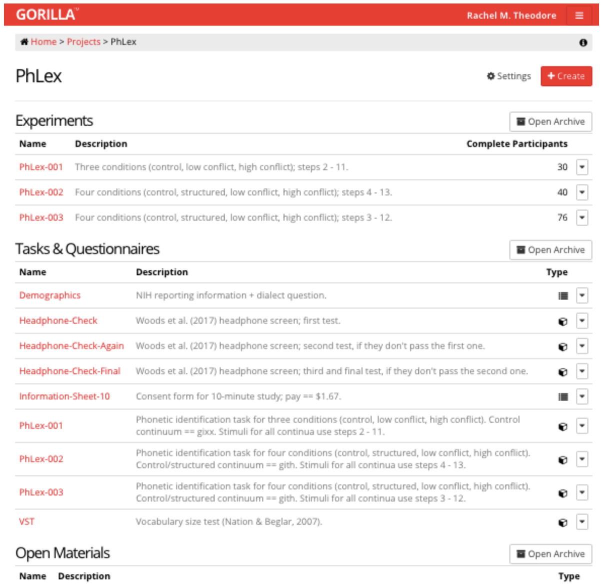
If you're listening to this talk, you've probably got a few reasons of your own for considering web-based studies. Let's talk about some tools. I'm going to focus on two, Gorilla Experiment Builder and Prolific. Other tools exist, but are outside of my area of expertise. To be honest, I landed on these two because I just couldn't figure out how to make MTurk work for me.

Gorilla is a both a lovely piece of software to build experiments - think of it as an equivalent to E-Prime, or PsychoPy, OpenSesame, or SuperLab. In addition to building experiments, it's also a server that can host your web-based studies. In my experience, there's no design that Gorilla can't handle.

It has extensive features including collaboration, version control, a repository of open materials consisting of experiments that others have designed, in addition to an engaging and comprehensive, multimedia support. It's not an exaggeration to say that an hour of your time would be sufficient to get you going in Gorilla.

In Gorilla, it's completely free to build experiments; the payment model is based on a token currency such that you're charged one token for each participant who completes your experiment.

- Projects consist of experiments, tasks & questionnaires, and open materials
- Experiments are sequenced tasks & questionnaires
- Open materials are publicly available tasks and questionnaires



PhLex

Experiments

Name	Description	Complete Participants
PhLex-001	Three conditions (control, low conflict, high conflict); steps 2 - 11.	30
PhLex-002	Four conditions (control, structured, low conflict, high conflict); steps 4 - 13.	40
PhLex-003	Four conditions (control, structured, low conflict, high conflict); steps 3 - 12.	76

Tasks & Questionnaires

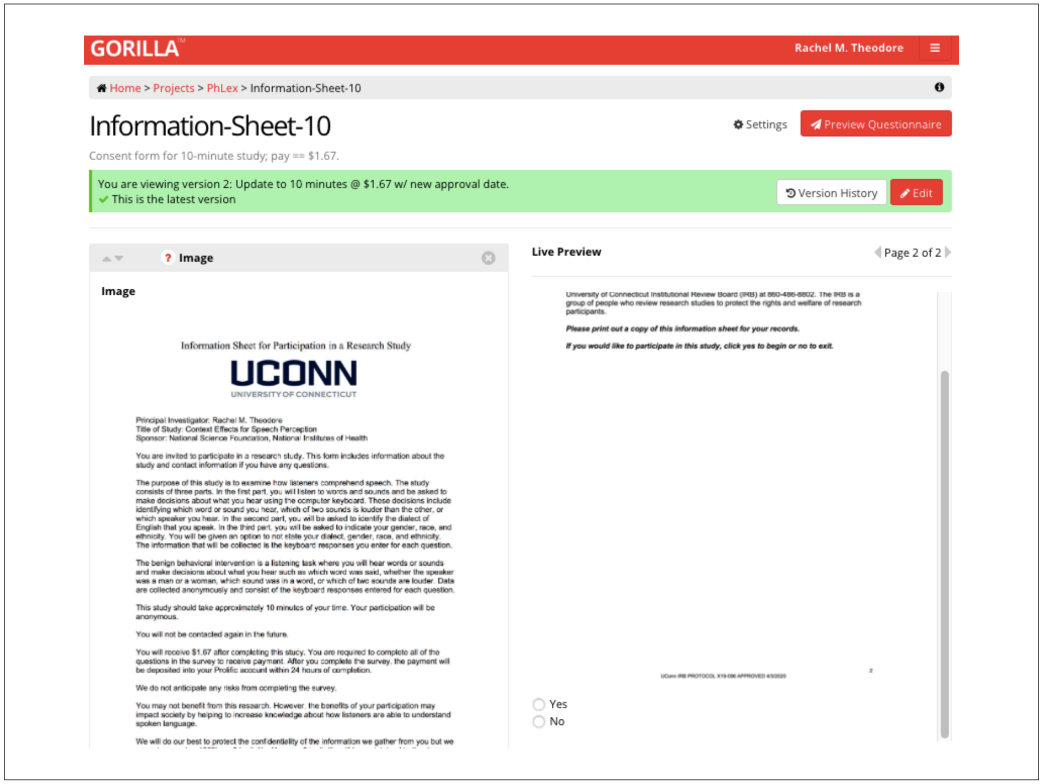
Name	Description	Type
Demographics	NIH reporting information + dialect question.	☰
Headphone-Check	Woods et al. (2017) headphone screen; first test.	⬢
Headphone-Check-Again	Woods et al. (2017) headphone screen; second test, if they don't pass the first one.	⬢
Headphone-Check-Final	Woods et al. (2017) headphone screen; third and final test, if they don't pass the second one.	⬢
Information-Sheet-10	Consent form for 10-minute study; pay == $1.67.	☰
PhLex-001	Phonetic identification task for three conditions (control, low conflict, high conflict). Control continuum == gixx. Stimuli for all continua use steps 2 - 11.	⬢
PhLex-002	Phonetic identification task for four conditions (control, structured, low conflict, high conflict). Control/structured continuum == gith. Stimuli for all continua use steps 4 - 13.	⬢
PhLex-003	Phonetic identification task for four conditions (control, structured, low conflict, high conflict). Control/structured continuum == gith. Stimuli for all continua use steps 3 - 12.	⬢
VST	Vocabulary size test (Nation & Beglar, 2007).	⬢

Open Materials

Name	Description	Type

Materials in Gorilla are organized around projects. Projects consist of experiments, tasks & questionnaires, and open materials. Experiments are sequenced tasks and questionnaires; that is, experiments are formed by combining the smaller bits into a sequenced order.

Open materials are any aspects of your project that you make publicly available on Gorilla.



Information-Sheet-10

Consent form for 10-minute study; pay == $1.67.

You are viewing version 2: Update to 10 minutes @ $1.67 w/ new approval date.
✓ This is the latest version

Image

Information Sheet for Participation in a Research Study

UCONN
UNIVERSITY OF CONNECTICUT

Principal Investigator: Rachel M. Theodore
Title of Study: Context Effects for Speech Perception
Sponsor: National Science Foundation, National Institutes of Health

You are invited to participate in a research study. This form includes information about the study and contact information if you have any questions.

The purpose of this study is to examine how listeners comprehend speech. The study consists of three parts. In the first part, you will listen to words and sounds and be asked to make decisions about what you hear using the computer keyboard. These decisions include identifying which word or sound you hear, which of two sounds is louder than the other, or which speaker you hear. In the second part, you will be asked to identify the dialect of English that you speak. In the third part, you will be asked to indicate your gender, race, and ethnicity. You will be given an option to not state your dialect, gender, race, and ethnicity. The information that will be collected is the keyboard responses you enter for each question.

The benign behavioral intervention is a listening task where you will hear words or sounds and make decisions about what you hear such as which word was said, whether the speaker was a man or a woman, which sound was in a word, or which of two sounds are louder. Data are collected anonymously and consist of the keyboard responses entered for each question.

This study should take approximately 10 minutes of your time. Your participation will be anonymous.

You will not be contacted again in the future.

You will receive $1.67 after completing this study. You are required to complete all of the questions in the survey to receive payment. After you complete the survey, the payment will be deposited into your Prolific account within 24 hours of completion.

We do not anticipate any risks from completing the survey.

You may not benefit from this research. However, the benefits of your participation may impact society by helping to increase knowledge about how listeners are able to understand spoken language.

We will do our best to protect the confidentiality of the information we gather from you but we

University of Connecticut Institutional Review Board (IRB) at 860-486-8802. The IRB is a group of people who review research studies to protect the rights and welfare of research participants.

Please print out a copy of this information sheet for your records.

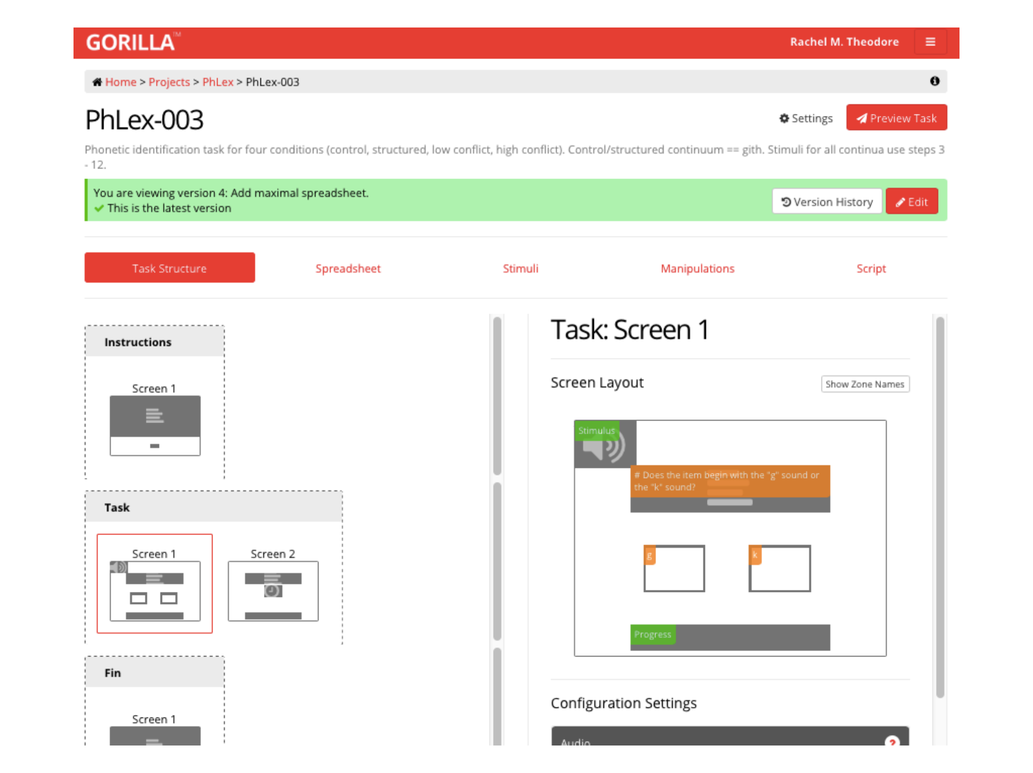
If you would like to participate in this study, click yes to begin or no to exit.

UConn IRB PROTOCOL X19-096 APPROVED 4/3/2020

2

○ Yes
○ No

Here's an example of the interface for making a questionnaire; specifically, this is how we built a consent form. It's two images, one for each page of the consent form, along with a response option.



PhLex-003

Phonetic identification task for four conditions (control, structured, low conflict, high conflict). Control/structured continuum == gith. Stimuli for all continua use steps 3 - 12.

You are viewing version 4: Add maximal spreadsheet.
✓ This is the latest version

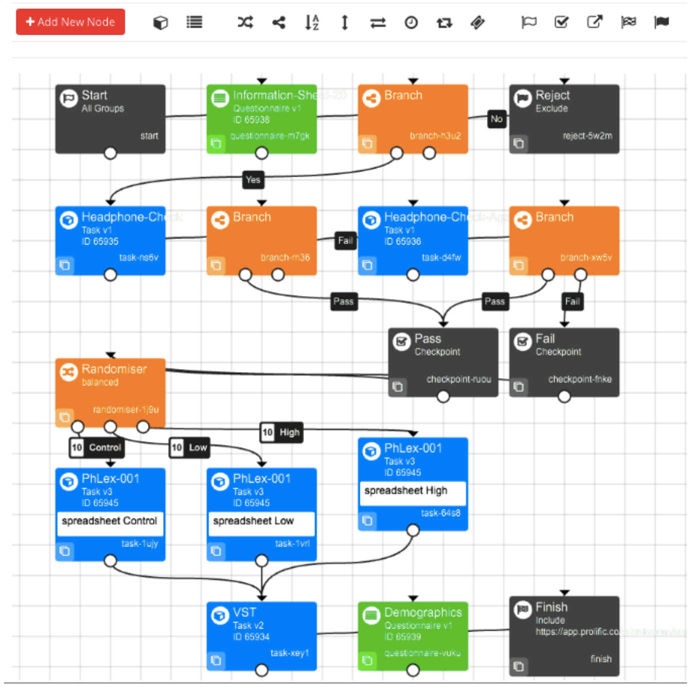
Task: Screen 1

Screen Layout

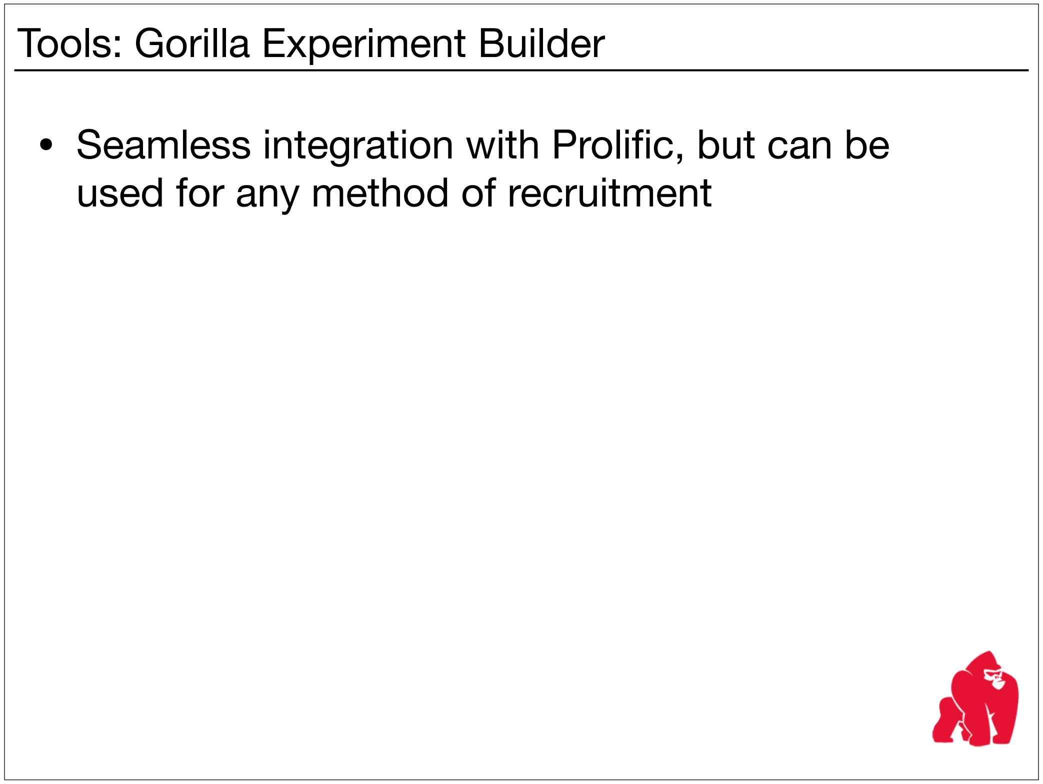
Does the item begin with the "g" sound or the "k" sound?

Configuration Settings

Here’s an example of the task building interface; it’s all super intuitive, does not require any programming knowledge, and is thoroughly described in the support documentation and videos.



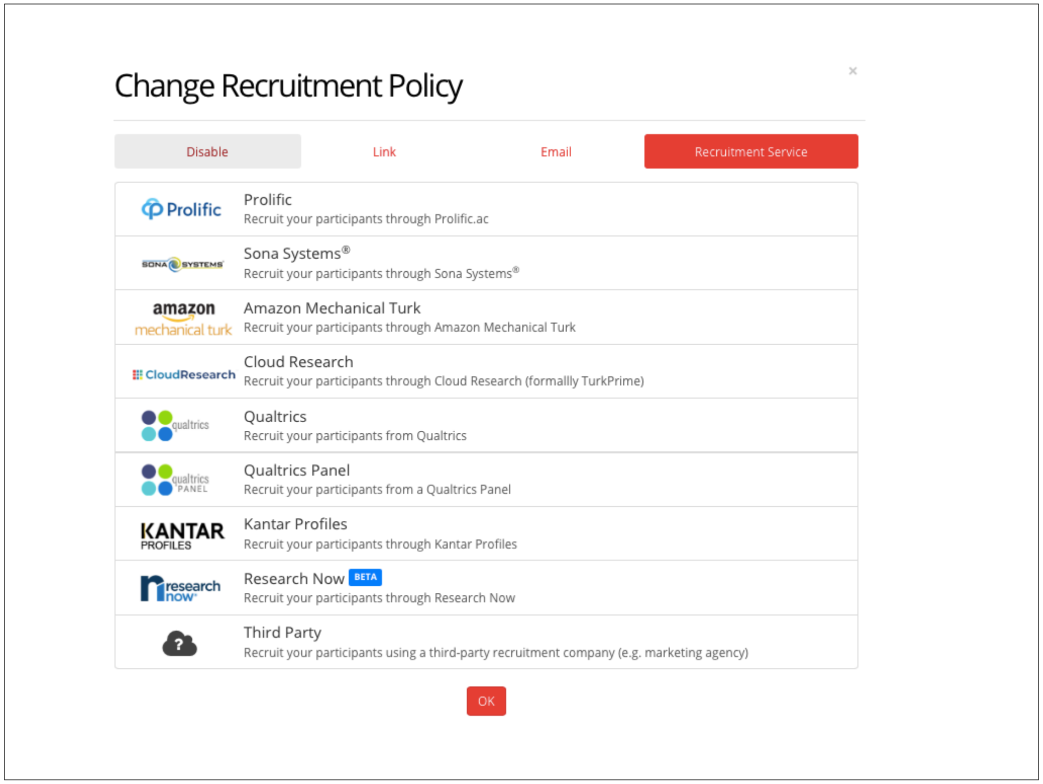
And here’s an example of the experiment interface. As you can see, the questionnaires, in green, and the tasks, in blue, have been ordered using logical branches, in orange. Here, people move from the start to the consent questionnaire, if they give consent, then they move to a headphone screen task, and if they pass, they then enter one of three experimental tasks. And so on —



Tools: Gorilla Experiment Builder

- Seamless integration with Prolific, but can be used for any method of recruitment

As I mentioned, Gorilla is not only an experiment builder, but it's also a server that hosts your web-based study. Gorilla integrates seamlessly with Prolific, a tool for recruiting participants that I'll talk about next - but you can use Gorilla with any method of recruitment.



Change Recruitment Policy

Prolific
Recruit your participants through Prolific.ac

Sona Systems®
Recruit your participants through Sona Systems®

Amazon Mechanical Turk
Recruit your participants through Amazon Mechanical Turk

Cloud Research
Recruit your participants through Cloud Research (formallly TurkPrime)

Qualtrics
Recruit your participants from Qualtrics

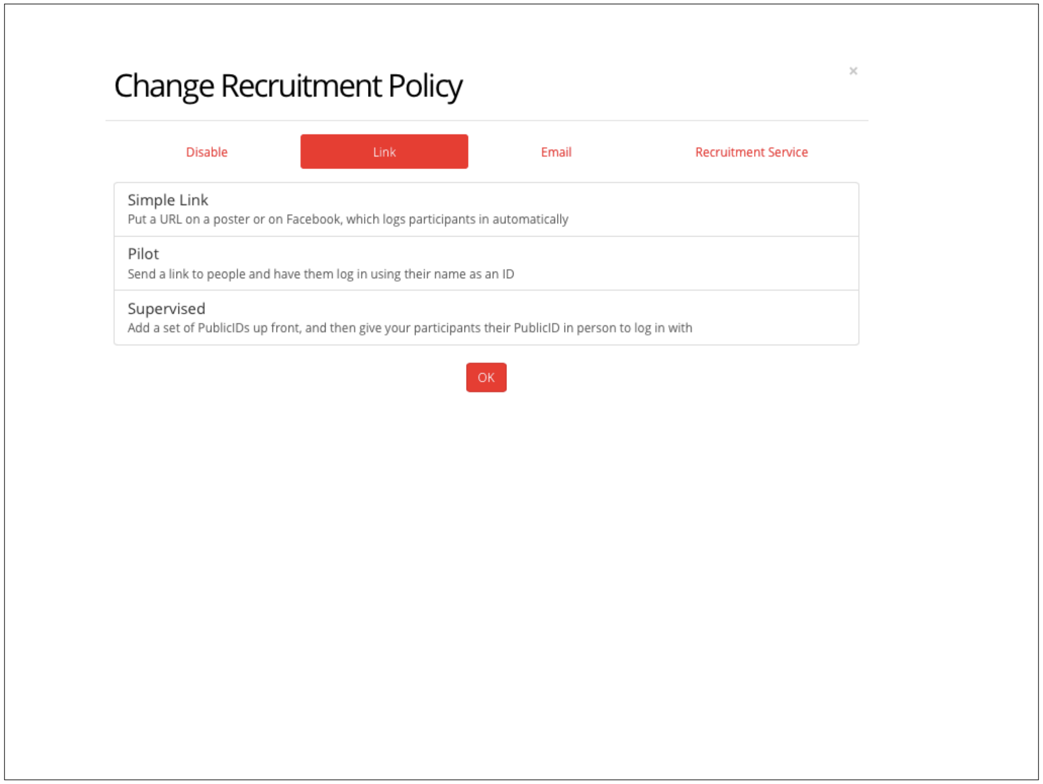
Qualtrics Panel
Recruit your participants from a Qualtrics Panel

Kantar Profiles
Recruit your participants through Kantar Profiles

Research Now BETA
Recruit your participants through Research Now

Third Party
Recruit your participants using a third-party recruitment company (e.g. marketing agency)

For example, Gorilla has built-in integration with numerous participant platforms including MTurk, Sona, and Qualtrics.



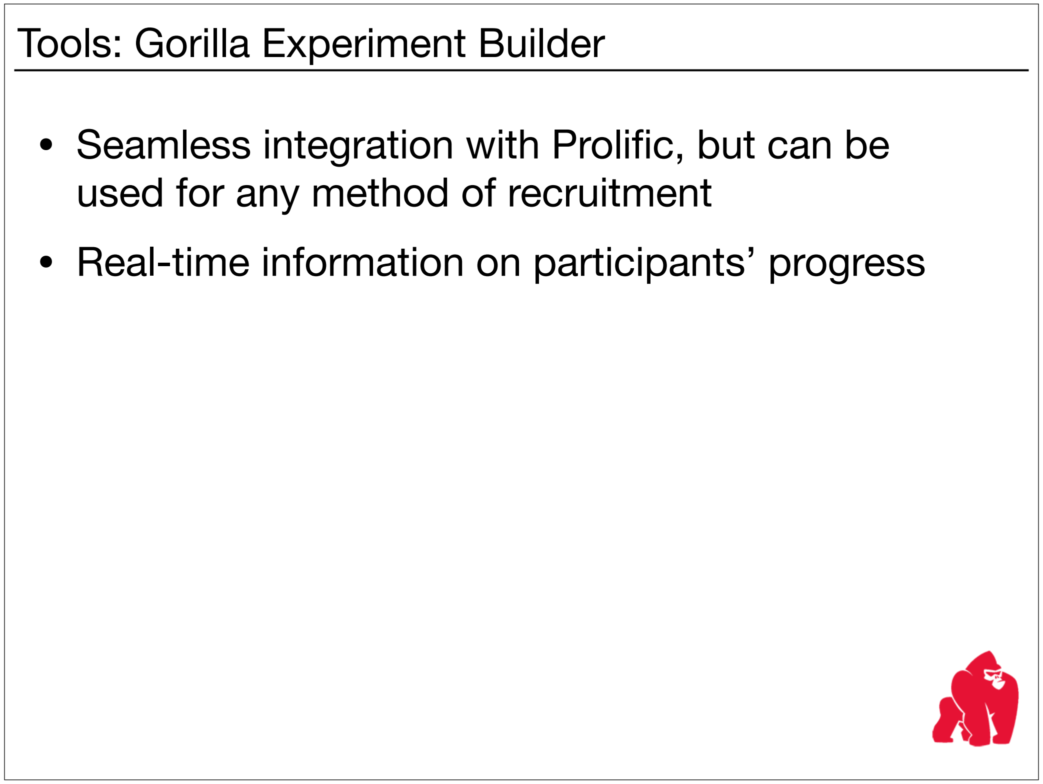
Change Recruitment Policy

Simple Link
Put a URL on a poster or on Facebook, which logs participants in automatically

Pilot
Send a link to people and have them log in using their name as an ID

Supervised
Add a set of PublicIDs up front, and then give your participants their PublicID in person to log in with

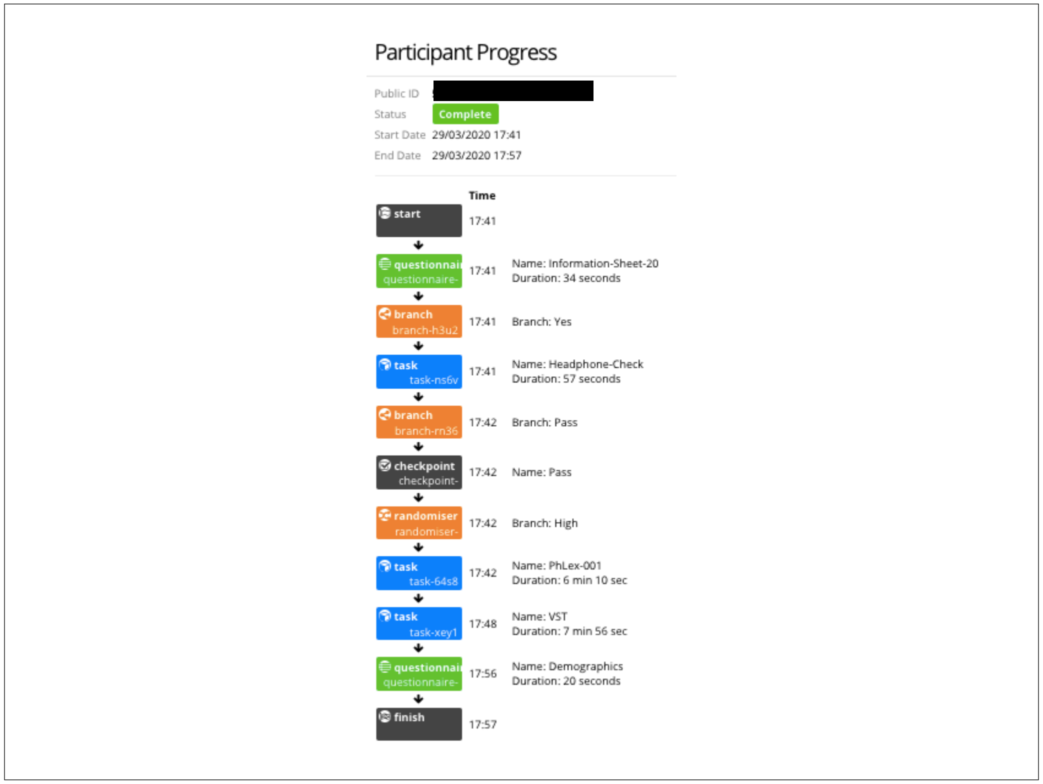
Or you could simply generate a link to distribute anywhere you wish. We do this for in-house pilot testing. For example, we could build an experiment, generate the link, and then drop it in the lab Slack for whoever is free to take the experiment for a run.



Tools: Gorilla Experiment Builder

- Seamless integration with Prolific, but can be used for any method of recruitment
- Real-time information on participants’ progress

In addition to providing lots of ways to get participants to your study, Gorilla also provides great visualization of real-time progress while people are completing your study.

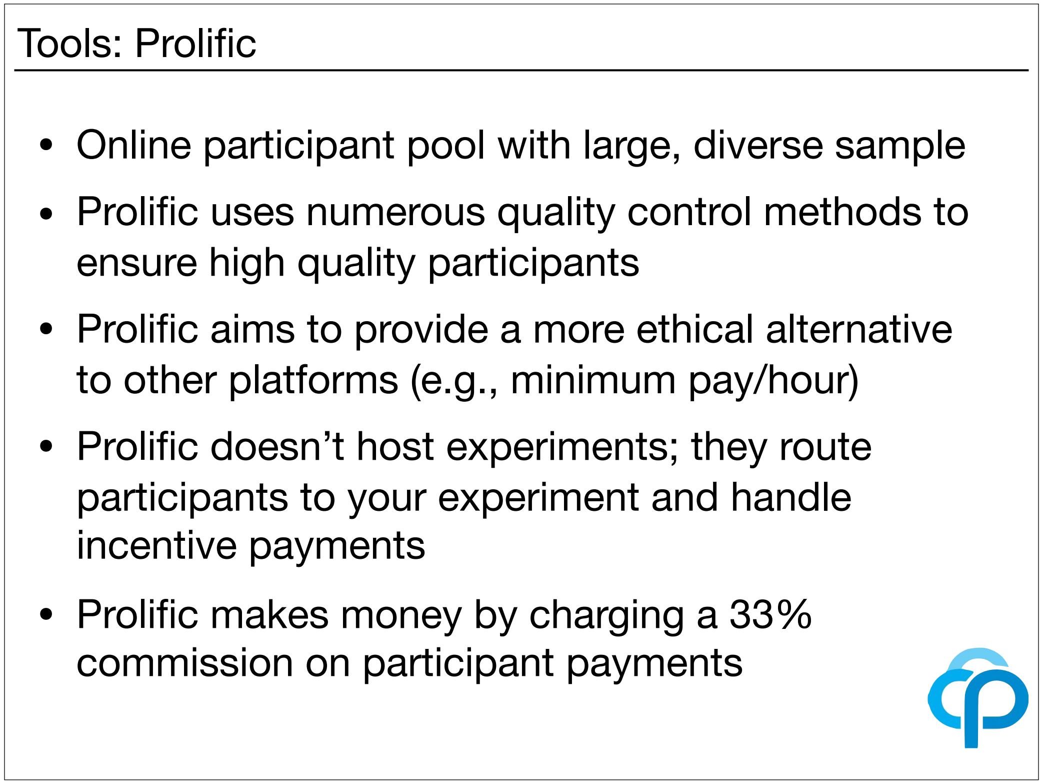


Participant Progress

Public ID [redacted]
Status Complete
Start Date 29/03/2020 17:41
End Date 29/03/2020 17:57

	Time	
start	17:41	
questionnai questionnaire-	17:41	Name: Information-Sheet-20 Duration: 34 seconds
branch branch-h3u2	17:41	Branch: Yes
task task-ns6v	17:41	Name: Headphone-Check Duration: 57 seconds
branch branch-rn36	17:42	Branch: Pass
checkpoint checkpoint-	17:42	Name: Pass
randomiser randomiser-	17:42	Branch: High
task task-64s8	17:42	Name: PhLex-001 Duration: 6 min 10 sec
task task-xey1	17:48	Name: VST Duration: 7 min 56 sec
questionnai questionnaire-	17:56	Name: Demographics Duration: 20 seconds
finish	17:57	

For example, you can watch participants in real time in terms of where they are in your experiment tree. I find this part especially satisfying to monitor — so much so that my trainees joke that it is my favorite television show.

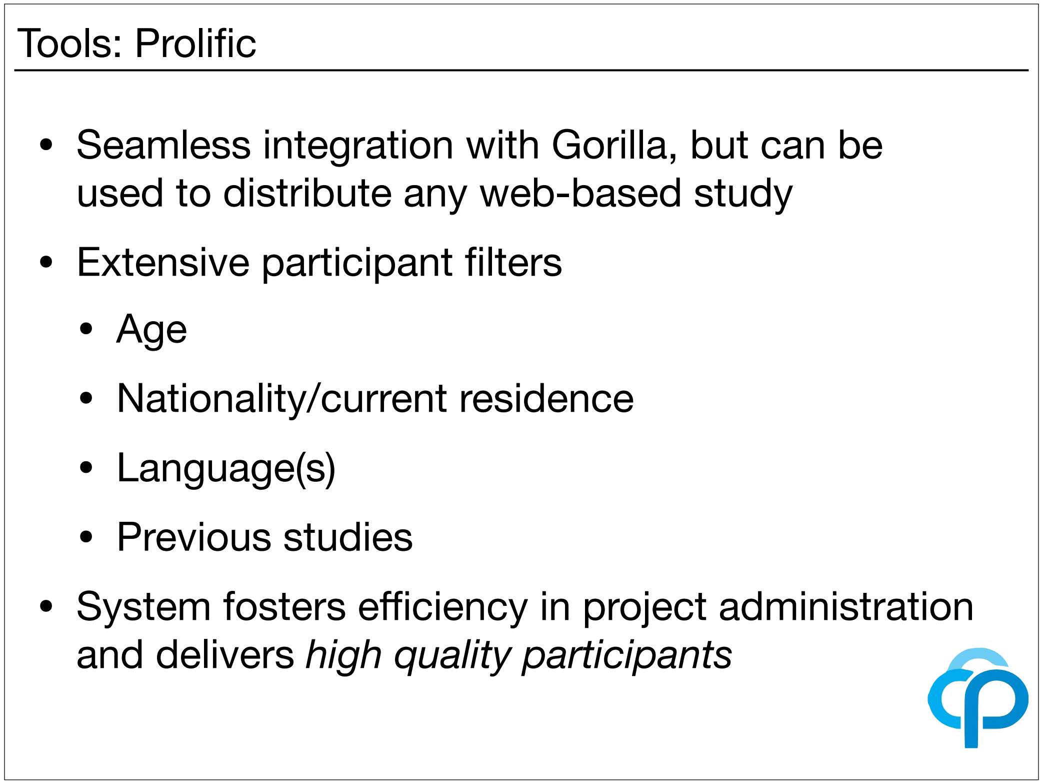


Tools: Prolific

- Online participant pool with large, diverse sample
- Prolific uses numerous quality control methods to ensure high quality participants
- Prolific aims to provide a more ethical alternative to other platforms (e.g., minimum pay/hour)
- Prolific doesn’t host experiments; they route participants to your experiment and handle incentive payments
- Prolific makes money by charging a 33% commission on participant payments

With Gorilla as the tool to build and host experiments, Prolific is a tool to find participants. This is an online pool with a large, diverse, sample. Anyone can sign up to join the pool!

Prolific does a host of things behind the scenes to promote high quality participants. They aim to provide a more ethical alternative to MTurk by setting a floor for participant incentives, among other researcher terms of service. Prolific doesn’t host the experiment — they are the middle men between your online study and participants. They make money by charging a 33% commission on participant payments — so if you give the participant $3.00, you’ll also give Prolific one dollar.

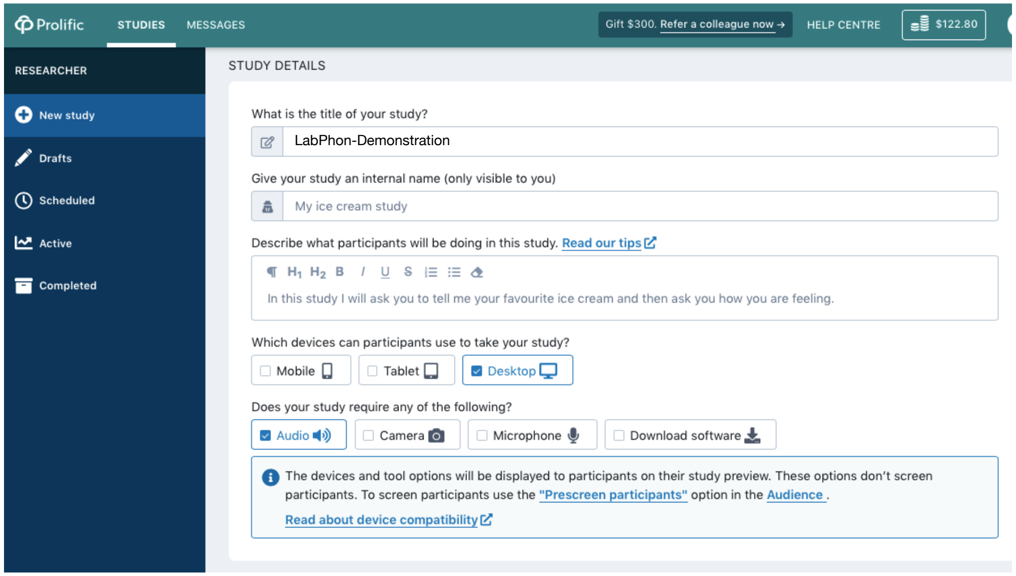


Tools: Prolific

- Seamless integration with Gorilla, but can be used to distribute any web-based study
- Extensive participant filters
  - Age
  - Nationality/current residence
  - Language(s)
  - Previous studies
- System fosters efficiency in project administration and delivers *high quality participants*

Prolific integrates seamlessly with Gorilla, but you can use Prolific to distribute any web-based study.

When people join the Prolific pool, they first they do is complete a series of over 150 questions that researchers can then use to filter who is recruited for their study. These include things like age, nationality, residence, language experience, and your own previous studies. Prolific is very receptive to adding new things to the on-boarding form as researchers indicate that new criteria are needed. The interface really streamlines project administration, including submitting receipts for reconciliation. Most importantly, the system excels at delivering high quality participants.



STUDY DETAILS

What is the title of your study?

LabPhon-Demonstration

Give your study an internal name (only visible to you)

My ice cream study

Describe what participants will be doing in this study. Read our tips

In this study I will ask you to tell me your favourite ice cream and then ask you how you are feeling.

Which devices can participants use to take your study?

□ Mobile □ Tablet ☑ Desktop

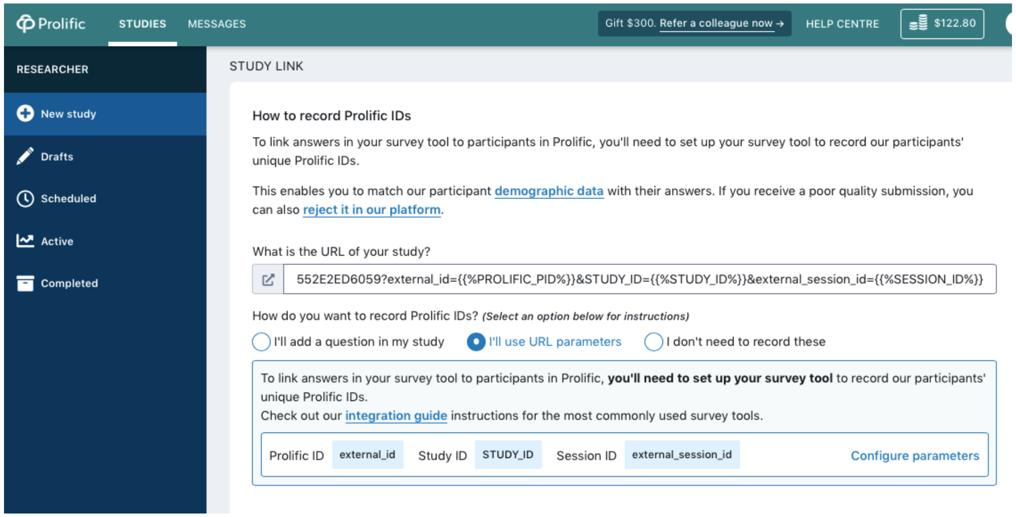
Does your study require any of the following?

☑ Audio □ Camera □ Microphone □ Download software

The devices and tool options will be displayed to participants on their study preview. These options don't screen participants. To screen participants use the "Prescreen participants" option in the Audience .

Read about device compatibility

Here's an example of the interface. I've indicated a study name; there's a place to provide an overview of the study that participants can see before they decide to do it; I've indicated some constraints, including that the study can't be completed on a tablet or mobile phone, and that there are audio stimuli.



STUDY LINK

How to record Prolific IDs

To link answers in your survey tool to participants in Prolific, you'll need to set up your survey tool to record our participants' unique Prolific IDs.

This enables you to match our participant demographic data with their answers. If you receive a poor quality submission, you can also reject it in our platform.

What is the URL of your study?

552E2ED6059?external_id={{%PROLIFIC_PID%}}&STUDY_ID={{%STUDY_ID%}}&external_session_id={{%SESSION_ID%}}

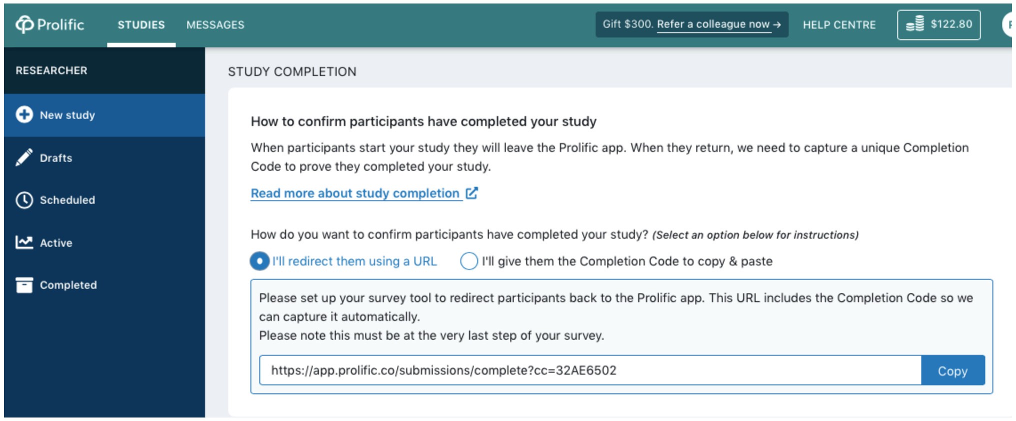
How do you want to record Prolific IDs? *(Select an option below for instructions)*

○ I'll add a question in my study ● I'll use URL parameters ○ I don't need to record these

To link answers in your survey tool to participants in Prolific, **you'll need to set up your survey tool** to record our participants' unique Prolific IDs.
Check out our integration guide instructions for the most commonly used survey tools.

Prolific ID external_id Study ID STUDY_ID Session ID external_session_id Configure parameters

You provide the link to your study — this is something that Gorilla generates for you.



STUDY COMPLETION

How to confirm participants have completed your study

When participants start your study they will leave the Prolific app. When they return, we need to capture a unique Completion Code to prove they completed your study.

Read more about study completion

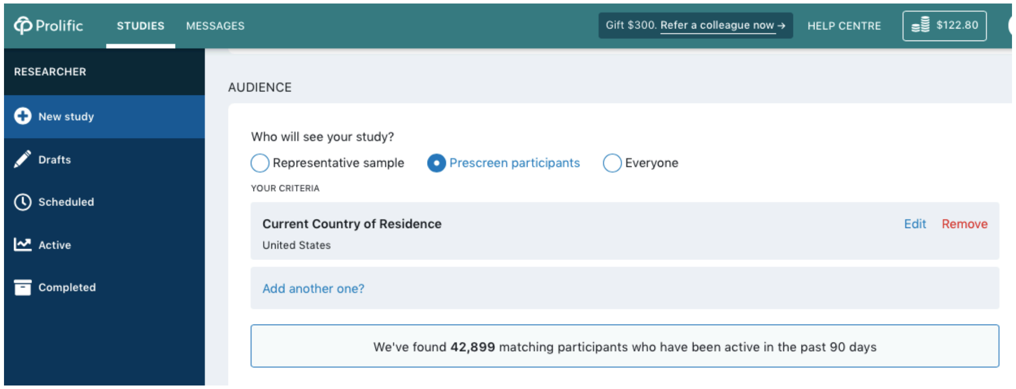
How do you want to confirm participants have completed your study? *(Select an option below for instructions)*

● I'll redirect them using a URL ○ I'll give them the Completion Code to copy & paste

Please set up your survey tool to redirect participants back to the Prolific app. This URL includes the Completion Code so we can capture it automatically.
Please note this must be at the very last step of your survey.

https://app.prolific.co/submissions/complete?cc=32AE6502

And Prolific gives you a link to add to the end of your Gorilla study so that participants are automatically routed back to Prolific for their incentive payment.



AUDIENCE

Who will see your study?

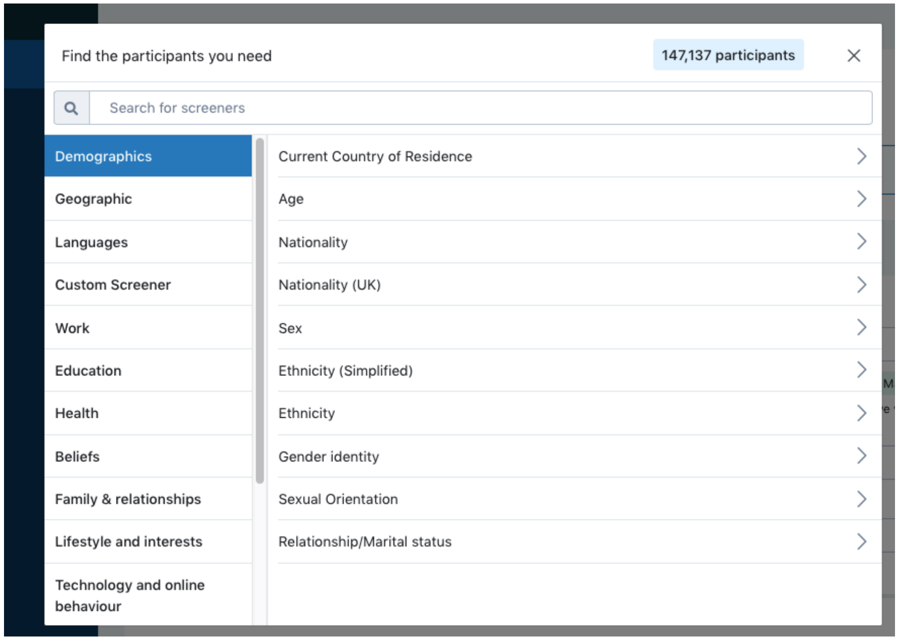
○ Representative sample ● Prescreen participants ○ Everyone

YOUR CRITERIA

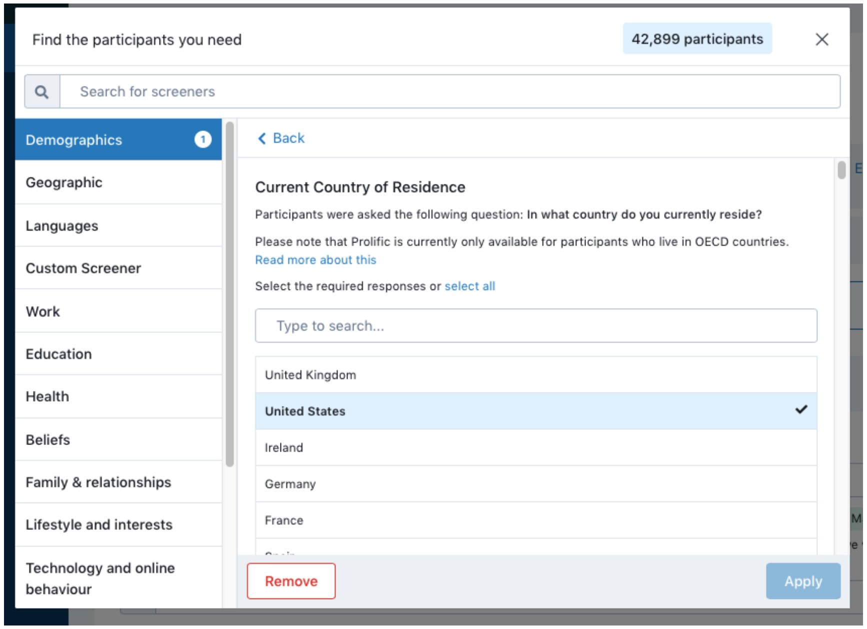
Current Country of Residence
United States

We've found **42,899** matching participants who have been active in the past 90 days

You get a real time display of how many active users meet your filter constraints —



And the interface for applying participant filters is very easy to use.



Current Country of Residence

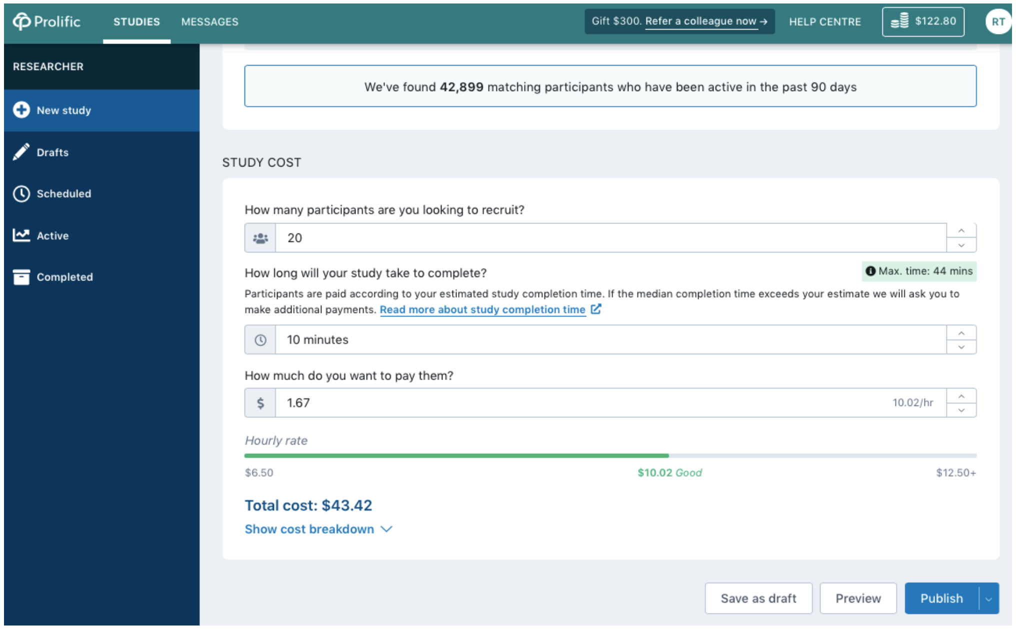
Participants were asked the following question: **In what country do you currently reside?**

Please note that Prolific is currently only available for participants who live in OECD countries.
Read more about this

Select the required responses or select all

Type to search...

- United Kingdom
- **United States** ✓
- Ireland
- Germany
- France



We've found **42,899** matching participants who have been active in the past 90 days

STUDY COST

How many participants are you looking to recruit?

20

How long will your study take to complete?

Max. time: 44 mins

Participants are paid according to your estimated study completion time. If the median completion time exceeds your estimate we will ask you to make additional payments. Read more about study completion time

10 minutes

How much do you want to pay them?

$ 1.67 10.02/hr

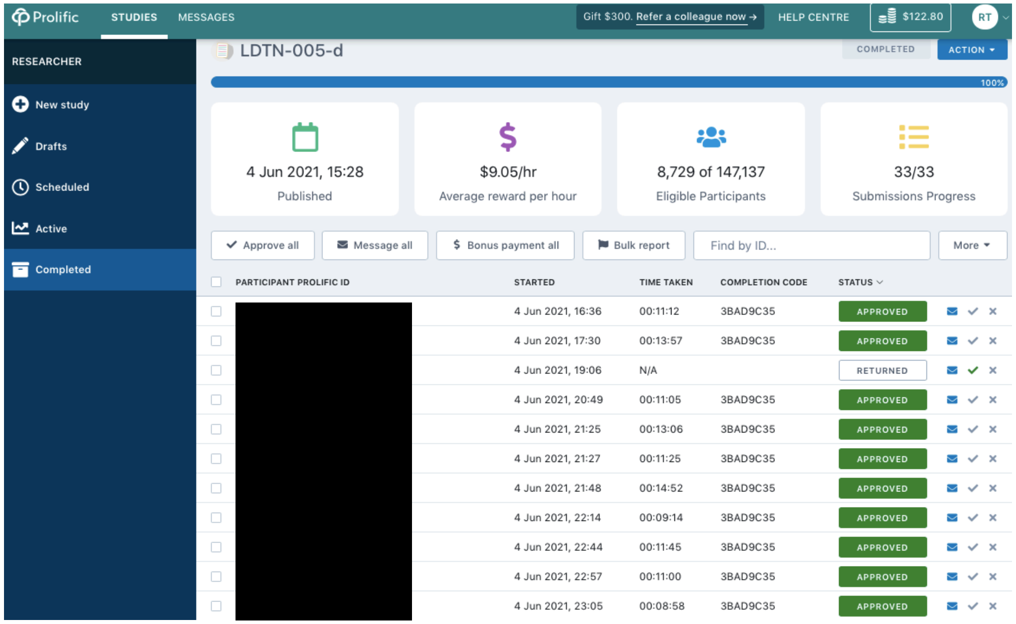
Hourly rate

$6.50 $10.02 *Good* $12.50+

Total cost: $43.42

Show cost breakdown

Everyone in a given study gets the same incentive payment, which is based on a good faith estimate of the average completion time.



LDTN-005-d

4 Jun 2021, 15:28
Published

$9.05/hr
Average reward per hour

8,729 of 147,137
Eligible Participants

33/33
Submissions Progress

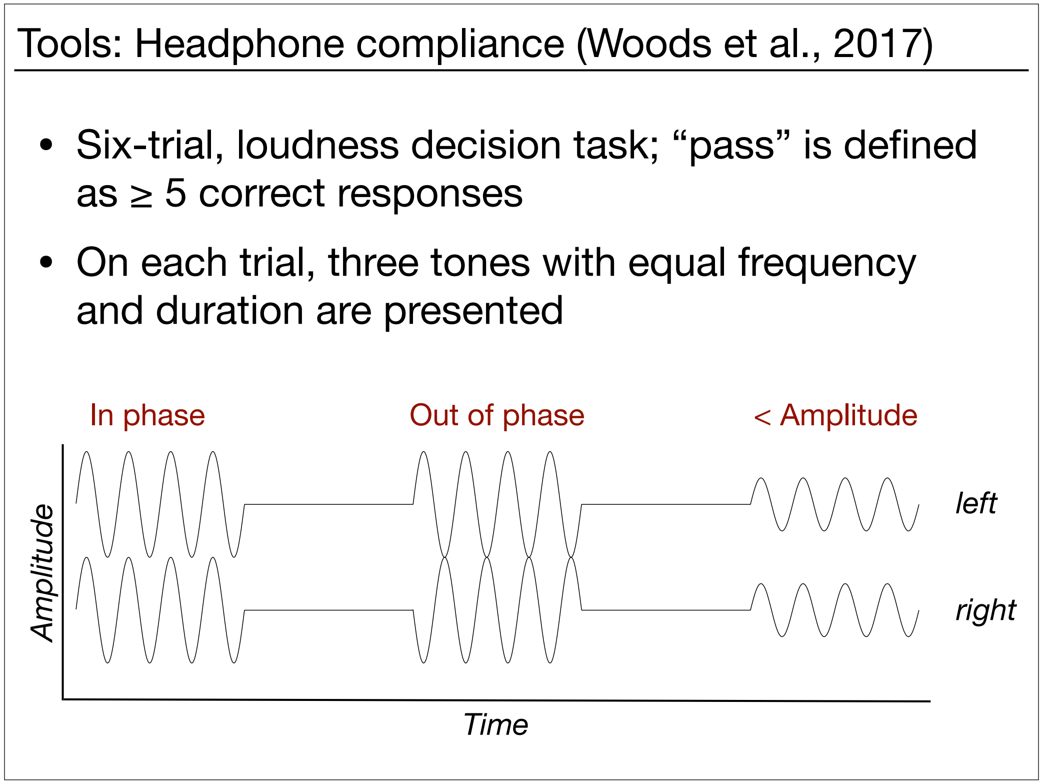
□	PARTICIPANT PROLIFIC ID	STARTED	TIME TAKEN	COMPLETION CODE	STATUS
□	[redacted]	4 Jun 2021, 16:36	00:11:12	3BAD9C35	APPROVED
□	[redacted]	4 Jun 2021, 17:30	00:13:57	3BAD9C35	APPROVED
□	[redacted]	4 Jun 2021, 19:06	N/A		RETURNED
□	[redacted]	4 Jun 2021, 20:49	00:11:05	3BAD9C35	APPROVED
□	[redacted]	4 Jun 2021, 21:25	00:13:06	3BAD9C35	APPROVED
□	[redacted]	4 Jun 2021, 21:27	00:11:25	3BAD9C35	APPROVED
□	[redacted]	4 Jun 2021, 21:48	00:14:52	3BAD9C35	APPROVED
□	[redacted]	4 Jun 2021, 22:14	00:09:14	3BAD9C35	APPROVED
□	[redacted]	4 Jun 2021, 22:44	00:11:45	3BAD9C35	APPROVED
□	[redacted]	4 Jun 2021, 22:57	00:11:00	3BAD9C35	APPROVED
□	[redacted]	4 Jun 2021, 23:05	00:08:58	3BAD9C35	APPROVED

The interface while a study is live, and also after it ends, is intuitive and informative.

Tools: Headphone compliance

- Woods et al. (2017)
- Milne et al. (2020)

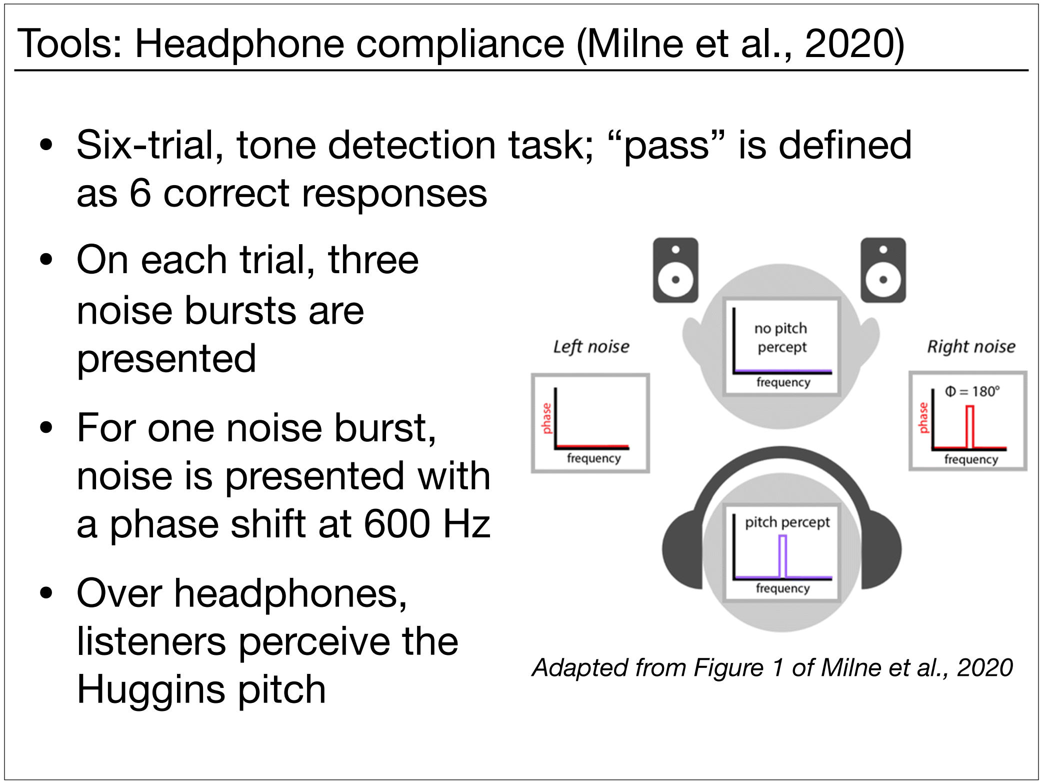
The last set of tools I’ll tell you about are two tasks designed to assess headphone use in web-based studies. Both of these are extremely clever, quick, dichotic listening tasks —



Tools: Headphone compliance (Woods et al., 2017)

- Six-trial, loudness decision task; “pass” is defined as ≥ 5 correct responses
- On each trial, three tones with equal frequency and duration are presented

In the interest of time, I won’t go into details, except to say that the Woods et al. task uses a dichotic phase cancelation manipulation to gauge headphone use from loudness judgments —

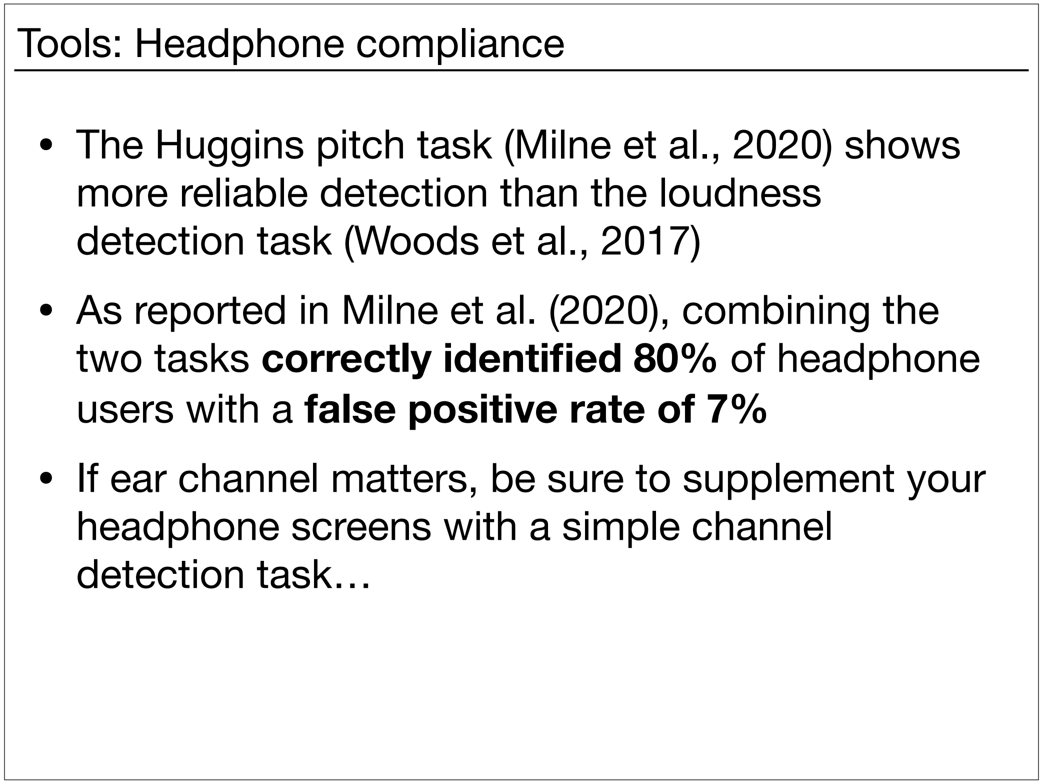


Tools: Headphone compliance (Milne et al., 2020)

- Six-trial, tone detection task; “pass” is defined as 6 correct responses
- On each trial, three noise bursts are presented
- For one noise burst, noise is presented with a phase shift at 600 Hz
- Over headphones, listeners perceive the Huggins pitch

Adapted from Figure 1 of Milne et al., 2020

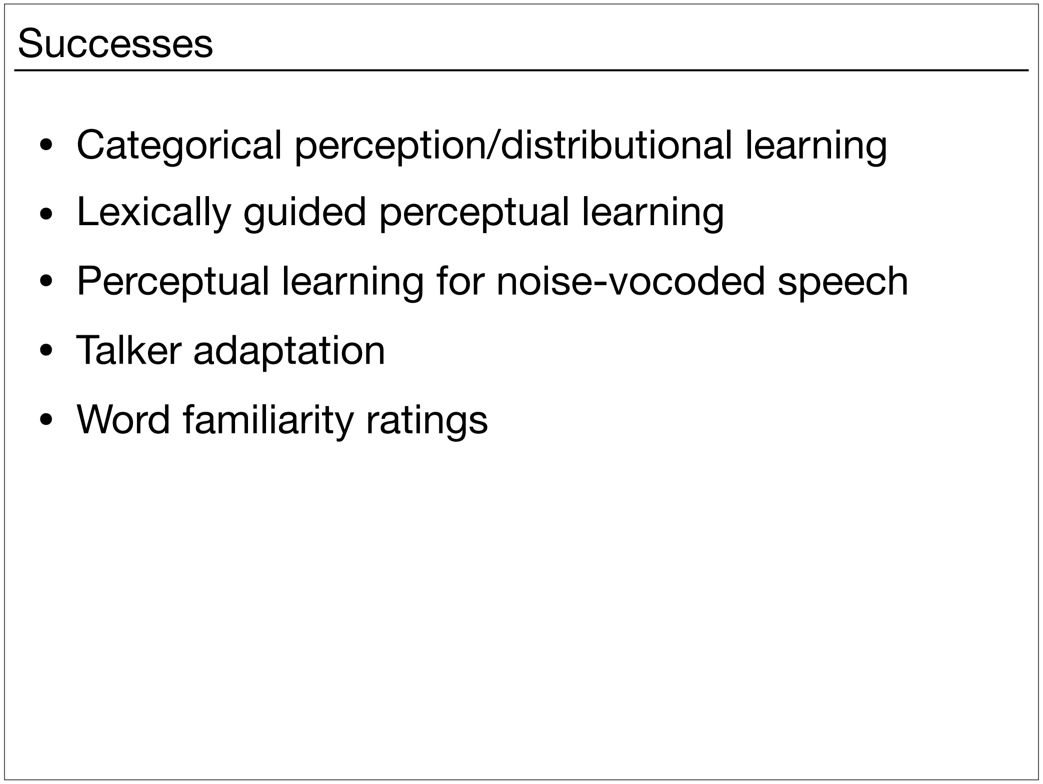
And the Milne et al. task is a Huggins pitch detection task —



Tools: Headphone compliance

- The Huggins pitch task (Milne et al., 2020) shows more reliable detection than the loudness detection task (Woods et al., 2017)
- As reported in Milne et al. (2020), combining the two tasks **correctly identified 80%** of headphone users with a **false positive rate of 7%**
- If ear channel matters, be sure to supplement your headphone screens with a simple channel detection task…

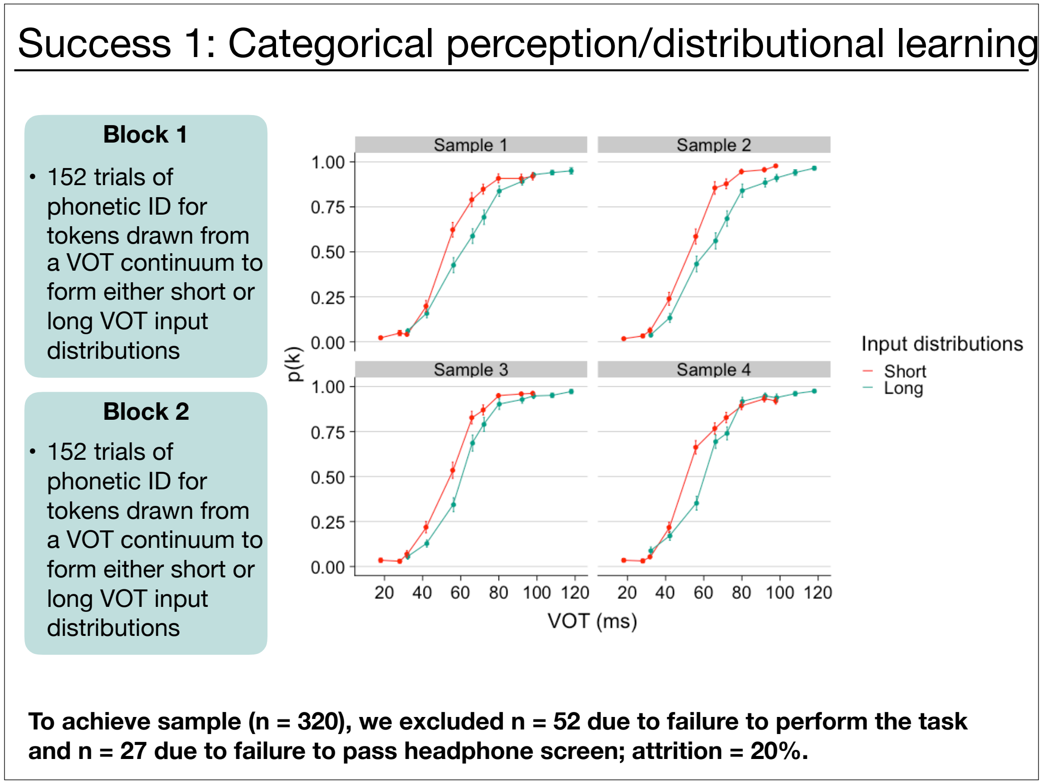
Combining the tasks only adds 12 total trials to your study and results in reasonable sensitivity and specificity in detecting the use of stereo headphones.



Successes

- Categorical perception/distributional learning
- Lexically guided perceptual learning
- Perceptual learning for noise-vocoded speech
- Talker adaptation
- Word familiarity ratings

In quick succession, I’m now going to share five successes we’ve had with web-based studies — selected to illustrate diversity in design and dependent measure. Preprints or postprints, along with data and analysis code, are available for all of these on our OSF repository.



Success 1: Categorical perception/distributional learning

Block 1

- 152 trials of phonetic ID for tokens drawn from a VOT continuum to form either short or long VOT input distributions

Block 2

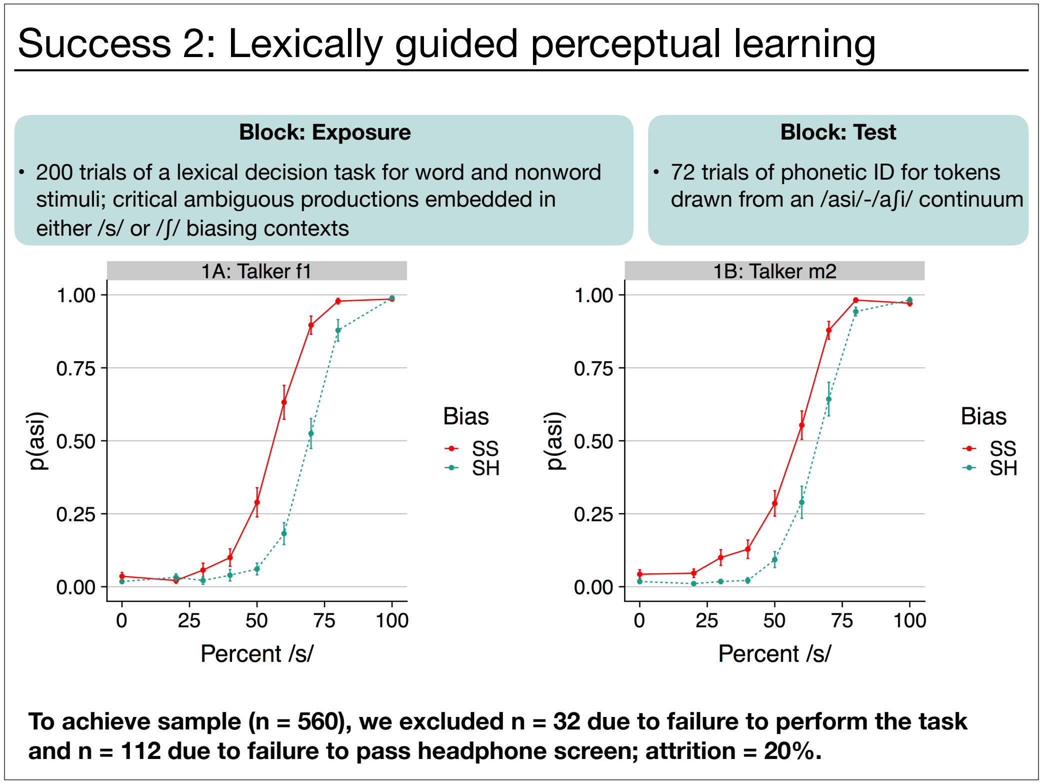
- 152 trials of phonetic ID for tokens drawn from a VOT continuum to form either short or long VOT input distributions

To achieve sample (n = 320), we excluded n = 52 due to failure to perform the task and n = 27 due to failure to pass headphone screen; attrition = 20%.

Success 1: Categorical perception and distributional learning.

In this study, four samples completed two blocks of a 2AFC category identification task. Across blocks, we manipulated the input distributions specifying the /g/ and /k/ categories. As you can see, all four samples showed the expected logistic relationship between category identification and VOT; critically, all four samples also yielded reliable evidence of distributional learning such that the identification function for the long VOT input distributions is displaced towards longer VOTs compared to the identification function for the short VOT input distributions.

Attrition due to failure to perform the task and failure to pass the headphone screen together yielded an attrition of 20%. As I go through these successes, you're going to see some variability in the attrition rate; in the tips and tricks section of this talk I'll share things we've learned to do to decrease our attrition rate.



Success 2: Lexically guided perceptual learning

Block: Exposure

- 200 trials of a lexical decision task for word and nonword stimuli; critical ambiguous productions embedded in either /s/ or /ʃ/ biasing contexts

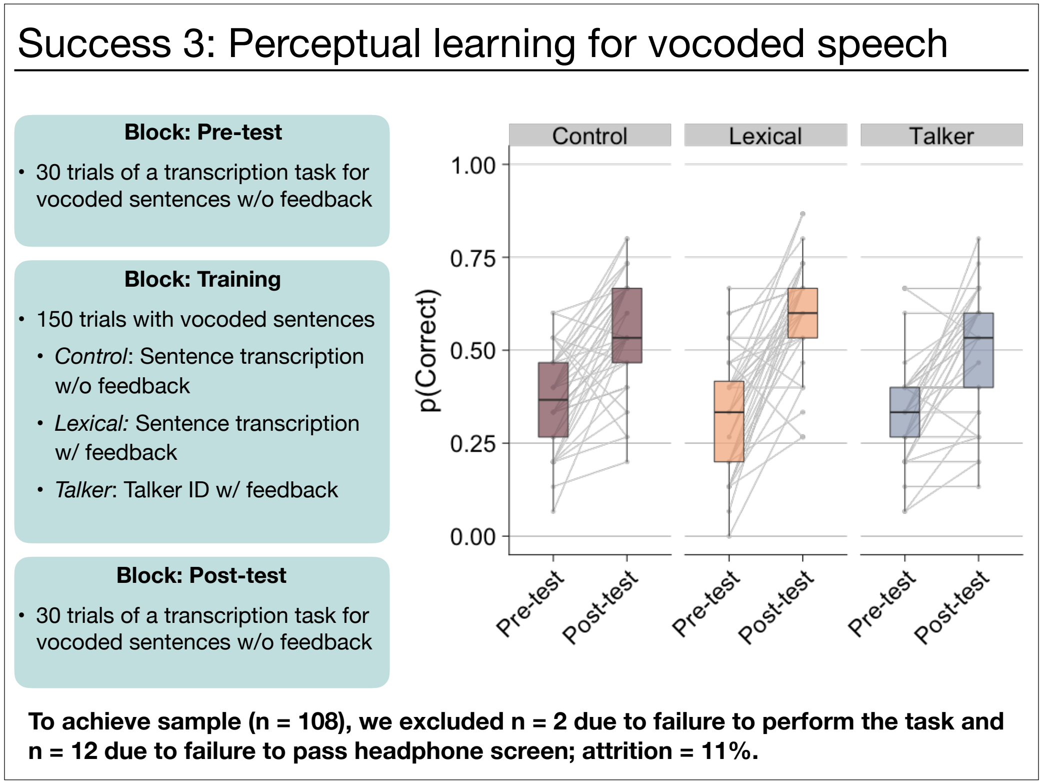
Block: Test

- 72 trials of phonetic ID for tokens drawn from an /asi/-/aʃi/ continuum

To achieve sample (n = 560), we excluded n = 32 due to failure to perform the task and n = 112 due to failure to pass headphone screen; attrition = 20%.

Success 2; lexically guided perceptual learning.

In this study, listeners completed two experimental blocks, an exposure phase and then a test phase. During exposure we manipulated the biasing lexical context for an ambiguous fricative. At test, listeners completed a 2AFC identification task for an *ashi* to *asi* continuum. Robust perceptual learning was observed for both tasks, with more *asi* responses at test for those biased to perceive the ambiguity as /s/ during exposure compared to those who were biased to perceive it as /ʃ/.



Success 3: Perceptual learning for vocoded speech

Block: Pre-test

- 30 trials of a transcription task for vocoded sentences w/o feedback

Block: Training

- 150 trials with vocoded sentences
  - *Control*: Sentence transcription w/o feedback
  - *Lexical:* Sentence transcription w/ feedback
  - *Talker*: Talker ID w/ feedback

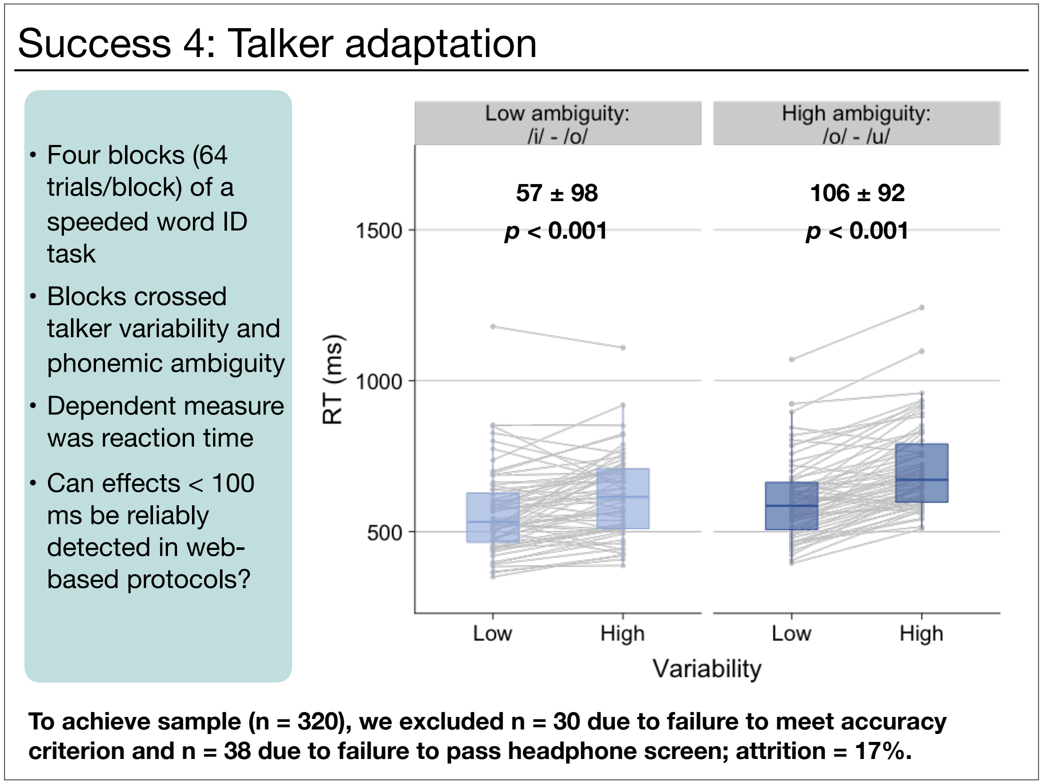
Block: Post-test

- 30 trials of a transcription task for vocoded sentences w/o feedback

To achieve sample (n = 108), we excluded n = 2 due to failure to perform the task and n = 12 due to failure to pass headphone screen; attrition = 11%.

Success 3; perceptual learning of noise-vocoded speech.

In this study, listeners completed pre-test, training, and post-test blocks. The task at pre- and post-test was free transcription of noise-vocoded sentences. Robust perceptual learning was observed, with transcription accuracy improved following training. Not shown here is a one-week follow-up test; web-based studies have truly opened logistical doors for us in terms of longitudinal experimental designs.



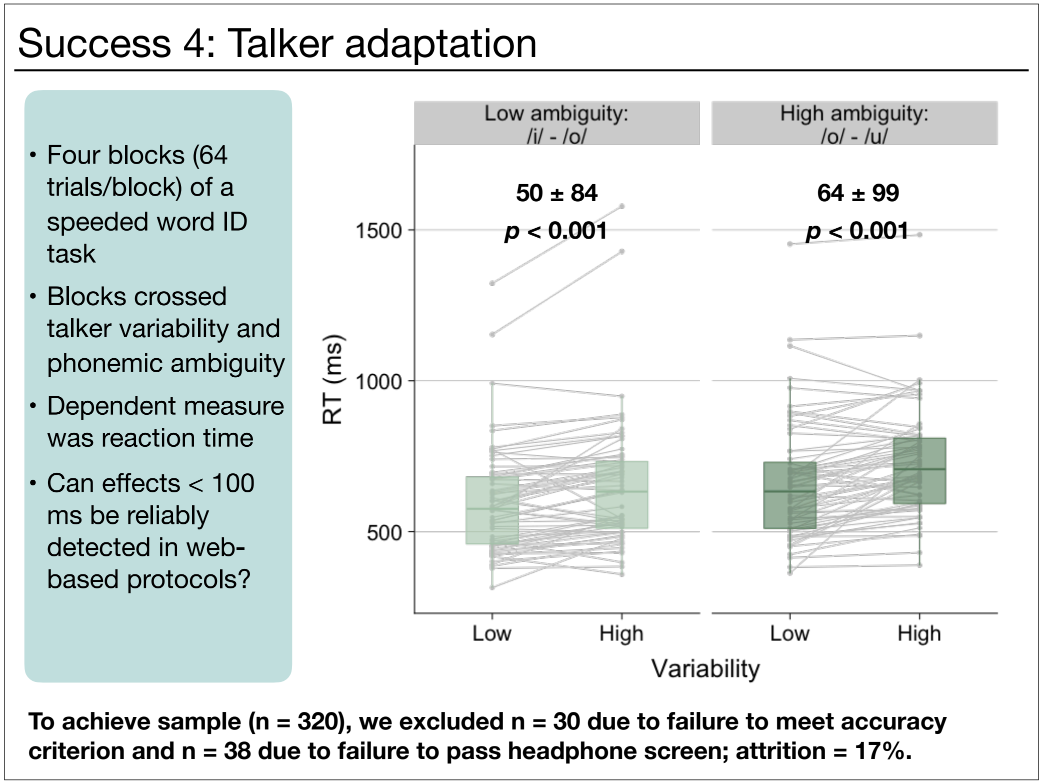
Success 4: Talker adaptation

- Four blocks (64 trials/block) of a speeded word ID task
- Blocks crossed talker variability and phonemic ambiguity
- Dependent measure was reaction time
- Can effects < 100 ms be reliably detected in web-based protocols?

To achieve sample (n = 320), we excluded n = 30 due to failure to meet accuracy criterion and n = 38 due to failure to pass headphone screen; attrition = 17%.

Success 4; talker adaptation.

In this study, listeners completed a speeded 2AFC word identification task for two blocks of stimuli, a single talker block, low variability, and a mixed talker block, high variability. This was our first foray into using RT as a dependent measure for a web-based design. As you can see, we had no challenges in reliably detecting variability effects under 100 ms in this sample —

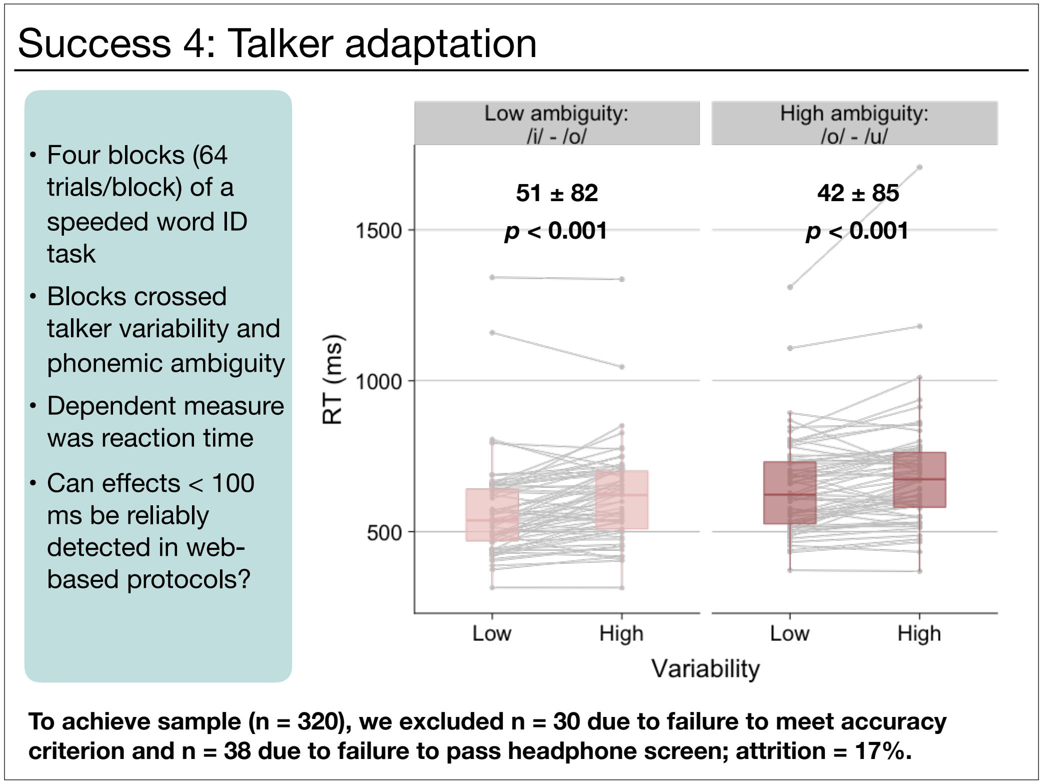


Success 4: Talker adaptation

- Four blocks (64 trials/block) of a speeded word ID task
- Blocks crossed talker variability and phonemic ambiguity
- Dependent measure was reaction time
- Can effects < 100 ms be reliably detected in web-based protocols?

To achieve sample (n = 320), we excluded n = 30 due to failure to meet accuracy criterion and n = 38 due to failure to pass headphone screen; attrition = 17%.

And in this sample.

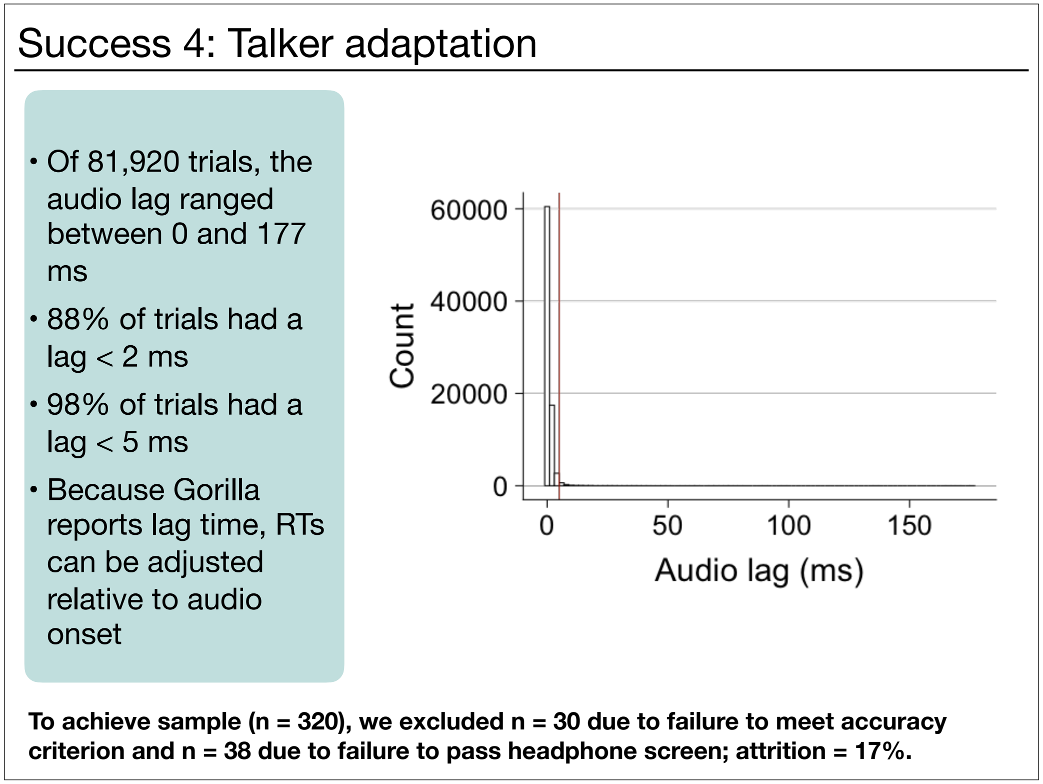


Success 4: Talker adaptation

- Four blocks (64 trials/block) of a speeded word ID task
- Blocks crossed talker variability and phonemic ambiguity
- Dependent measure was reaction time
- Can effects < 100 ms be reliably detected in web-based protocols?

To achieve sample (n = 320), we excluded n = 30 due to failure to meet accuracy criterion and n = 38 due to failure to pass headphone screen; attrition = 17%.

And in this sample.



Success 4: Talker adaptation

- Of 81,920 trials, the audio lag ranged between 0 and 177 ms
- 88% of trials had a lag < 2 ms
- 98% of trials had a lag < 5 ms
- Because Gorilla reports lag time, RTs can be adjusted relative to audio onset

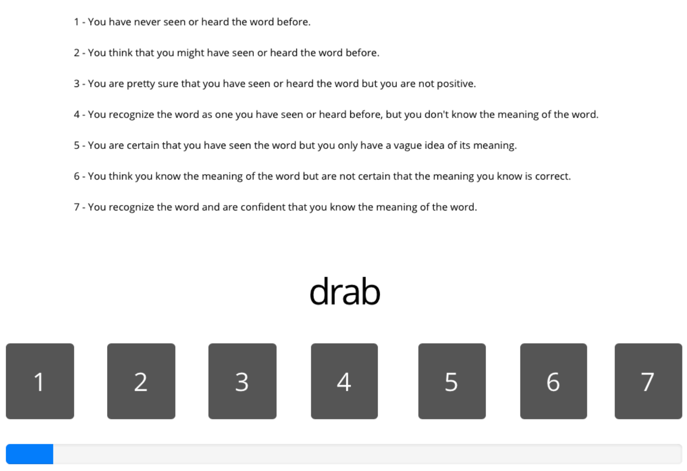
To achieve sample (n = 320), we excluded n = 30 due to failure to meet accuracy criterion and n = 38 due to failure to pass headphone screen; attrition = 17%.

Gorilla does a lot of things in the background to optimize stimulus presentation and response timing. In your data file, you get not only the timing of a button response, but also the lag between when the audio stimulus was set to play and when it actually did, which might vary based on a participant's particular system. Because of this, you can correct your RTs to reflect the actual onset of stimulus presentation. We analyzed the lag across all trials for 320 participants in this study and it was exquisite; 98% of the trials had a lag less than 5 milliseconds.

Overall for this study, the magnitude of effects, standard deviations of effects, and proportion of RT outliers were incredibly similar to in-lab work with similar paradigms.

Success 5: Word familiarity ratings

- Web-based administration of the Word Familiarity Test (WordFAM)
- Rating task for 150 items; 50 items from each of 3 frequency categories
- Adapted from Lewellen et al. (1993) and Pisoni (2007)



1 - You have never seen or heard the word before.

2 - You think that you might have seen or heard the word before.

3 - You are pretty sure that you have seen or heard the word but you are not positive.

4 - You recognize the word as one you have seen or heard before, but you don't know the meaning of the word.

5 - You are certain that you have seen the word but you only have a vague idea of its meaning.

6 - You think you know the meaning of the word but are not certain that the meaning you know is correct.

7 - You recognize the word and are confident that you know the meaning of the word.

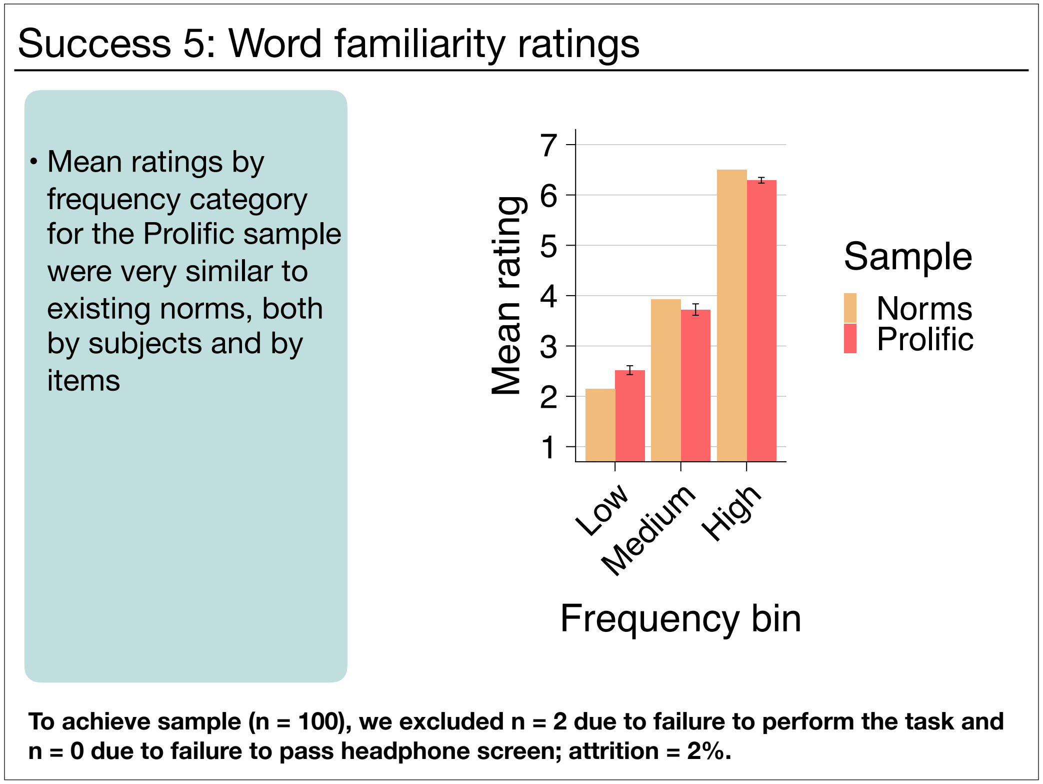
drab

To achieve sample (n = 100), we excluded n = 2 due to failure to perform the task and n = 0 due to failure to pass headphone screen; attrition = 2%.

Success 5; word familiarity ratings.

One thing that I think is especially frightening when moving to web-based studies, especially if you’re using the Prolific pool, is that you can’t see or interact with your participants. As a consequence, researchers often fear that they aren’t who they say they are.

To try and develop a tool that might help researchers verify some aspect of language competence, such as, are they a native English speaker as they say they are, we ported a paper-and-pencil vocabulary assessment to Gorilla. This is the word familiarity test developed by David Pisoni and colleagues. On each of 150 trials, participants see a word and are asked to indicate their familiarity with this word.

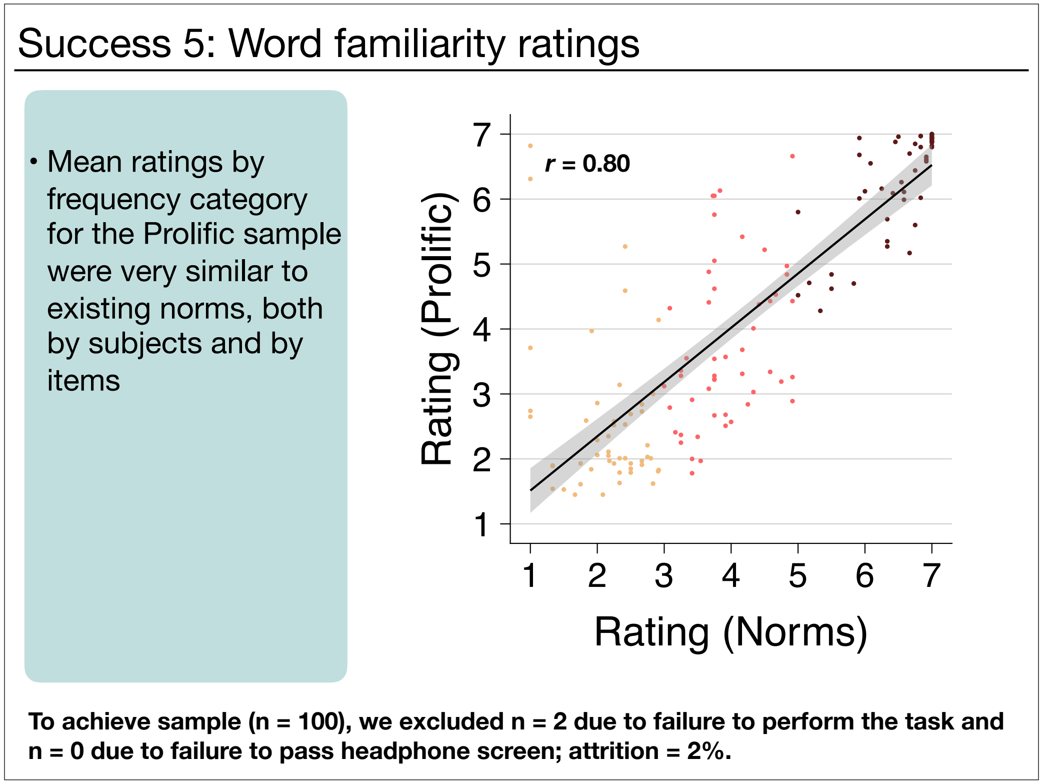


Success 5: Word familiarity ratings

- Mean ratings by frequency category for the Prolific sample were very similar to existing norms, both by subjects and by items

To achieve sample (n = 100), we excluded n = 2 due to failure to perform the task and n = 0 due to failure to pass headphone screen; attrition = 2%.

The 150 items represent 50 items in each of three frequency bins. David had normative data for this assessment so we were able to compare the mean ratings for the Prolific sample to the existing norms of the in-lab Hoosier participants. And look at that — mean ratings were *incredibly* similar between the two samples.

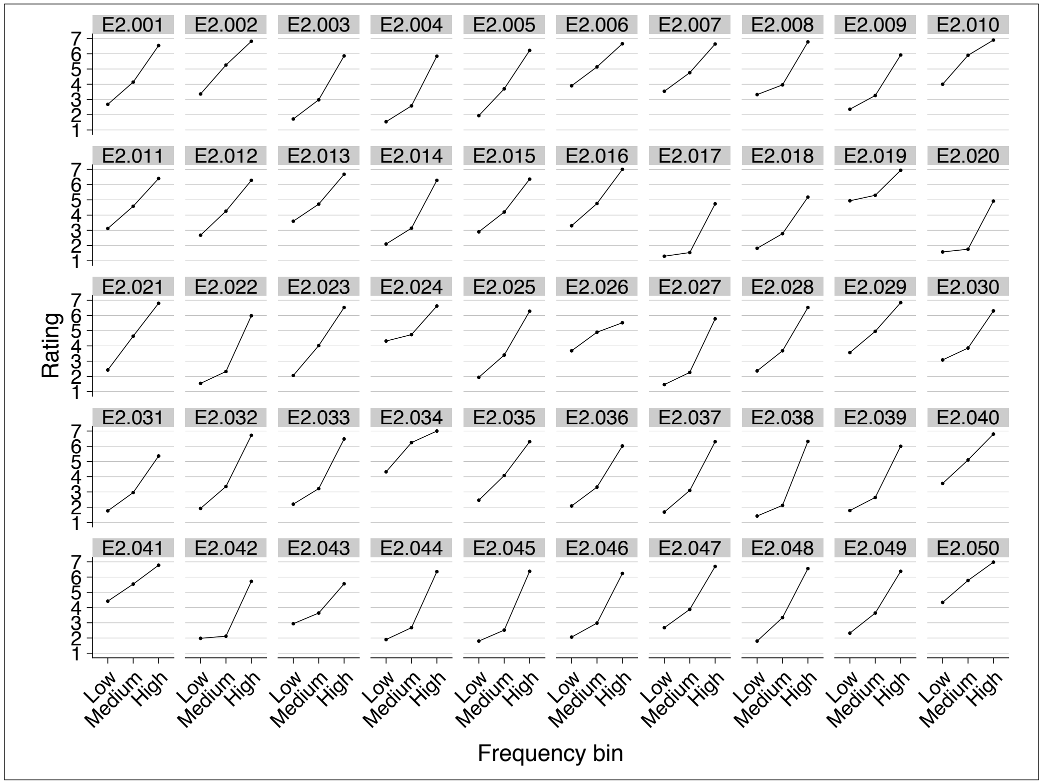


Success 5: Word familiarity ratings

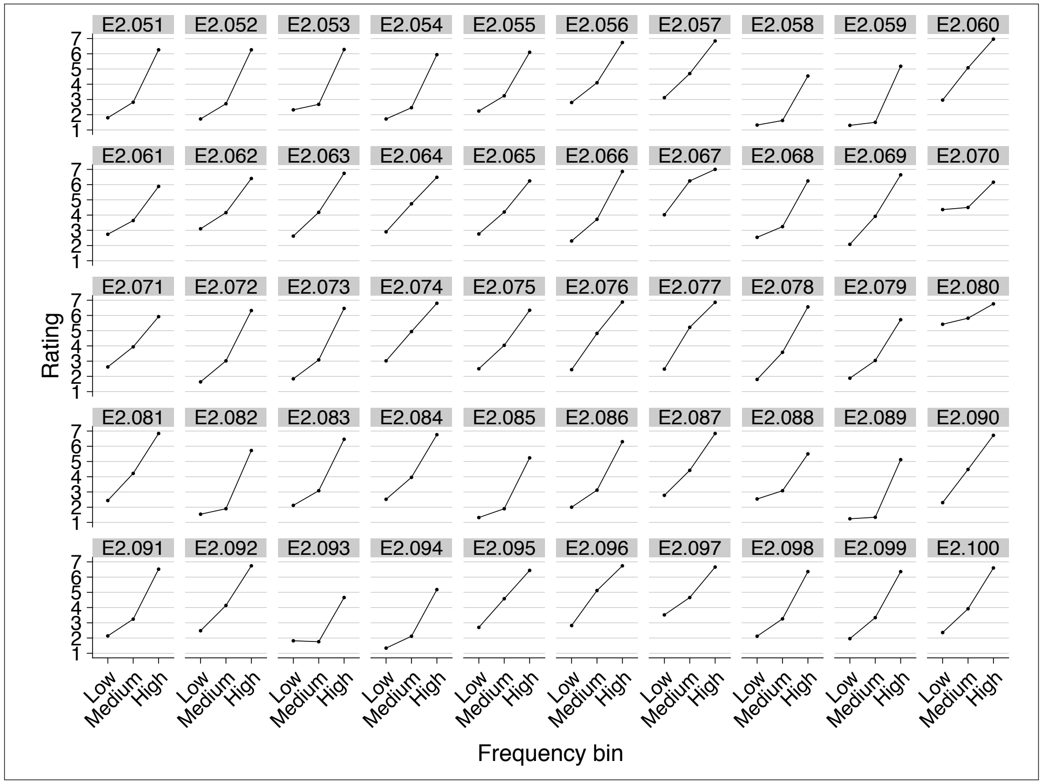
- Mean ratings by frequency category for the Prolific sample were very similar to existing norms, both by subjects and by items

To achieve sample (n = 100), we excluded n = 2 due to failure to perform the task and n = 0 due to failure to pass headphone screen; attrition = 2%.

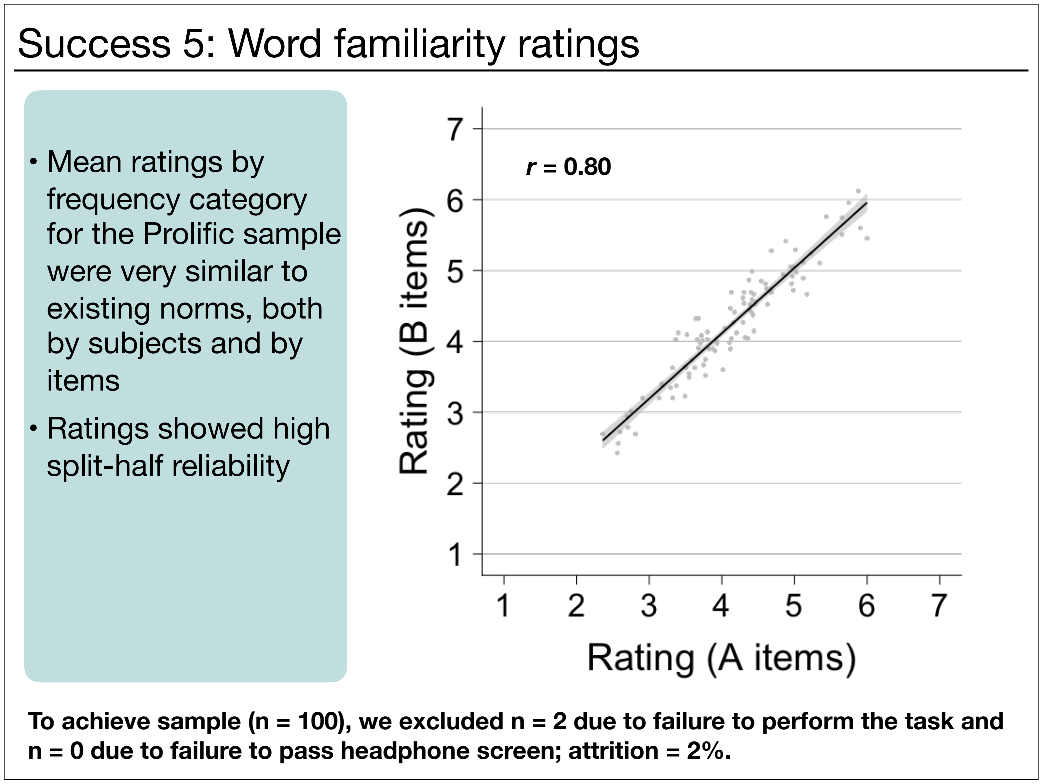
Not only were norms similar across samples by subjects, but they also tracked closely by items.



Here are individual subject functions; all subjects show the expected frequency effect. One also sees robust, and sensible, individual variation. For example, E2.019 shows overall higher ratings than E2.020, but both show the expected frequency effect.



Here are the other 50 people in the sample; lovely individual patterns as well.

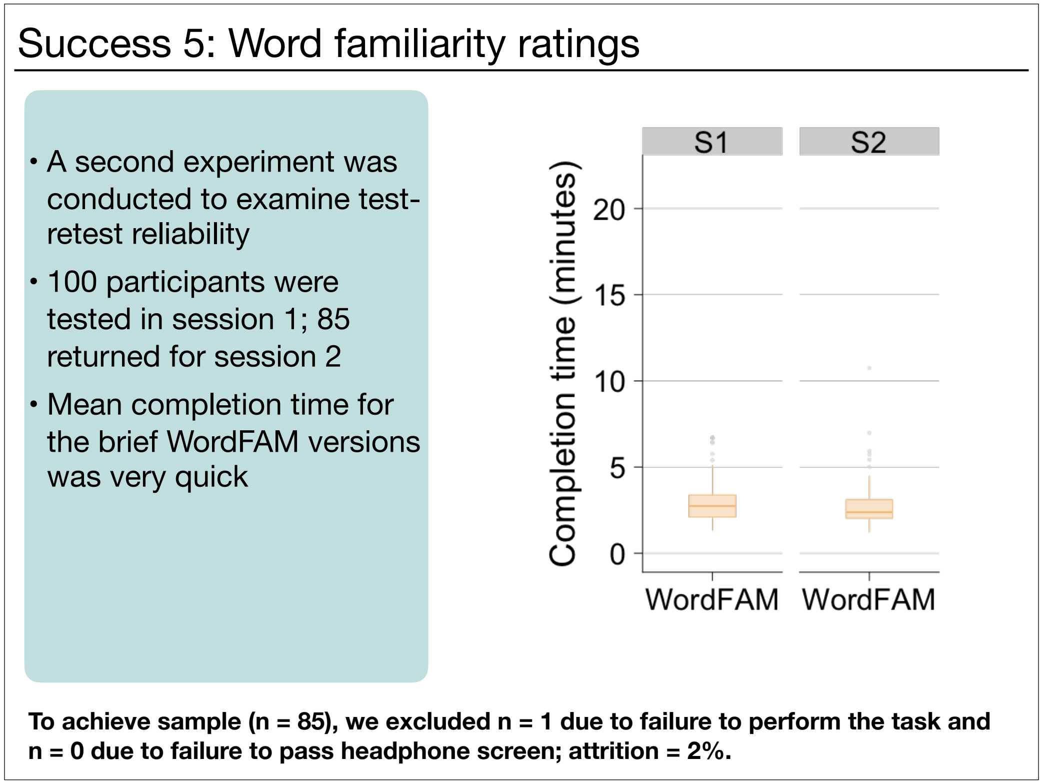


Success 5: Word familiarity ratings

- Mean ratings by frequency category for the Prolific sample were very similar to existing norms, both by subjects and by items
- Ratings showed high split-half reliability

To achieve sample (n = 100), we excluded n = 2 due to failure to perform the task and n = 0 due to failure to pass headphone screen; attrition = 2%.

In this first sample of 100 participants, we observed incredibly high split-half reliability, which we used as motivation to try and develop an even briefer assessment.

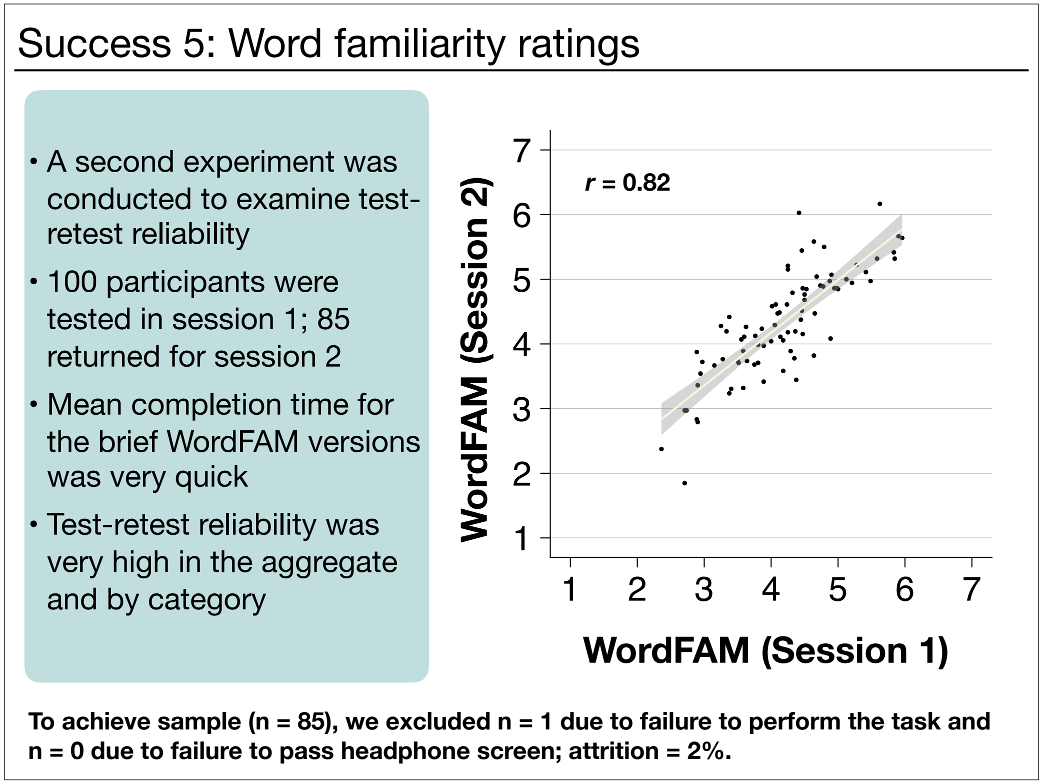


Success 5: Word familiarity ratings

- A second experiment was conducted to examine test-retest reliability
- 100 participants were tested in session 1; 85 returned for session 2
- Mean completion time for the brief WordFAM versions was very quick

To achieve sample (n = 85), we excluded n = 1 due to failure to perform the task and n = 0 due to failure to pass headphone screen; attrition = 2%.

Specifically, a second experiment was conducted that included 85 participants who completed two brief versions of the WordFam test, separated by about two weeks in time. Mean completion time was around four minutes —

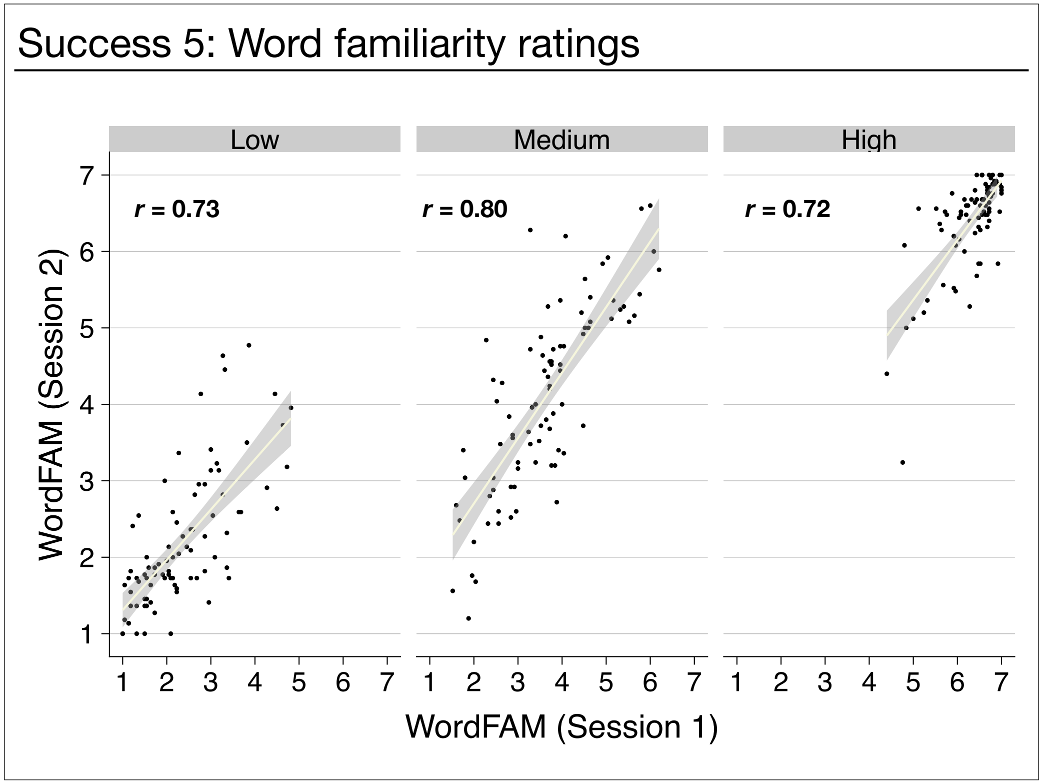


Success 5: Word familiarity ratings

- A second experiment was conducted to examine test-retest reliability
- 100 participants were tested in session 1; 85 returned for session 2
- Mean completion time for the brief WordFAM versions was very quick
- Test-retest reliability was very high in the aggregate and by category

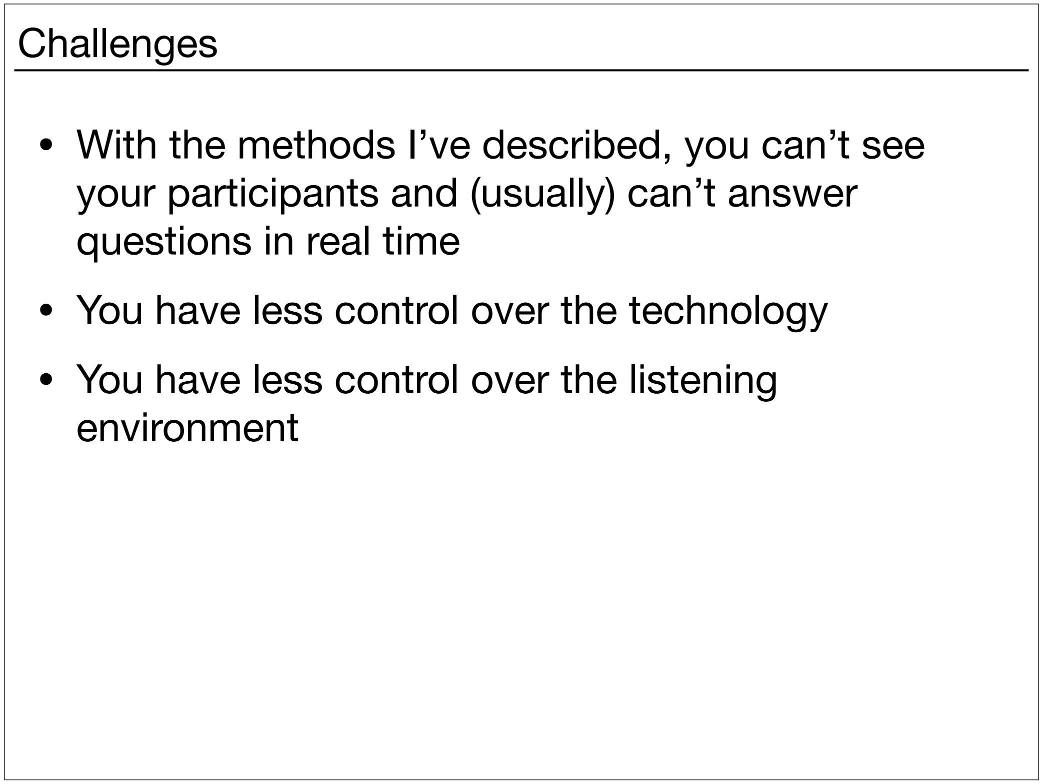
To achieve sample (n = 85), we excluded n = 1 due to failure to perform the task and n = 0 due to failure to pass headphone screen; attrition = 2%.

Test-retest reliability was incredibly high in the aggregate —



Success 5: Word familiarity ratings

And within each frequency category. These assessments, along with a second measure of vocabulary knowledge, will soon be available as Open Materials on Gorilla.



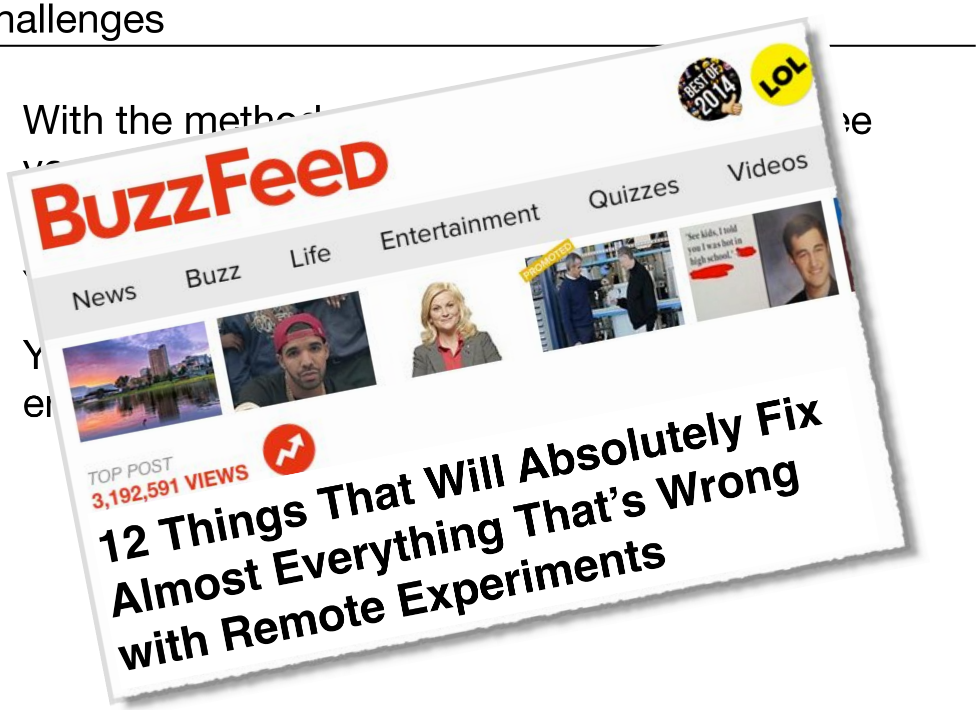
Challenges

- With the methods I’ve described, you can’t see your participants and (usually) can’t answer questions in real time
- You have less control over the technology
- You have less control over the listening environment

These five success are just a sample of what’s possible — what a time to be alive. That being said, it’s also true that web-based speech perception studies are not without challenges given the loss of some control over both participant’s specific hardware and the listening environment.

Challenges

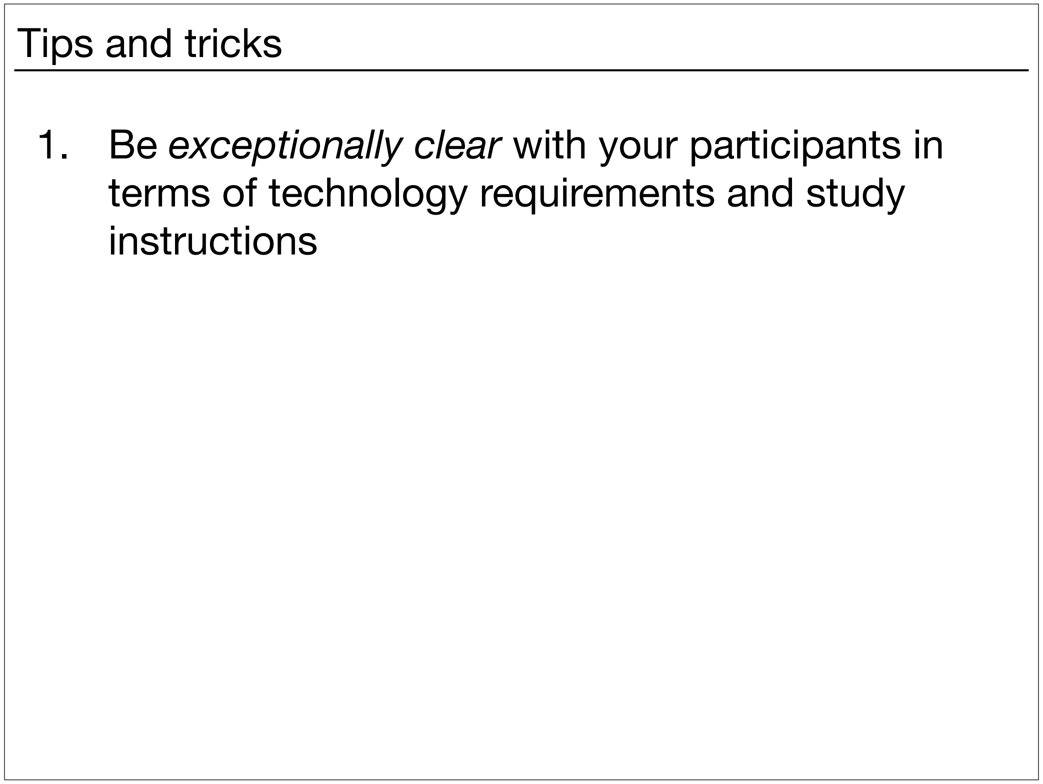
- With the metho ee
  yo
-
- Y
  er



TOP POST
3,192,591 VIEWS

12 Things That Will Absolutely Fix Almost Everything That's Wrong with Remote Experiments

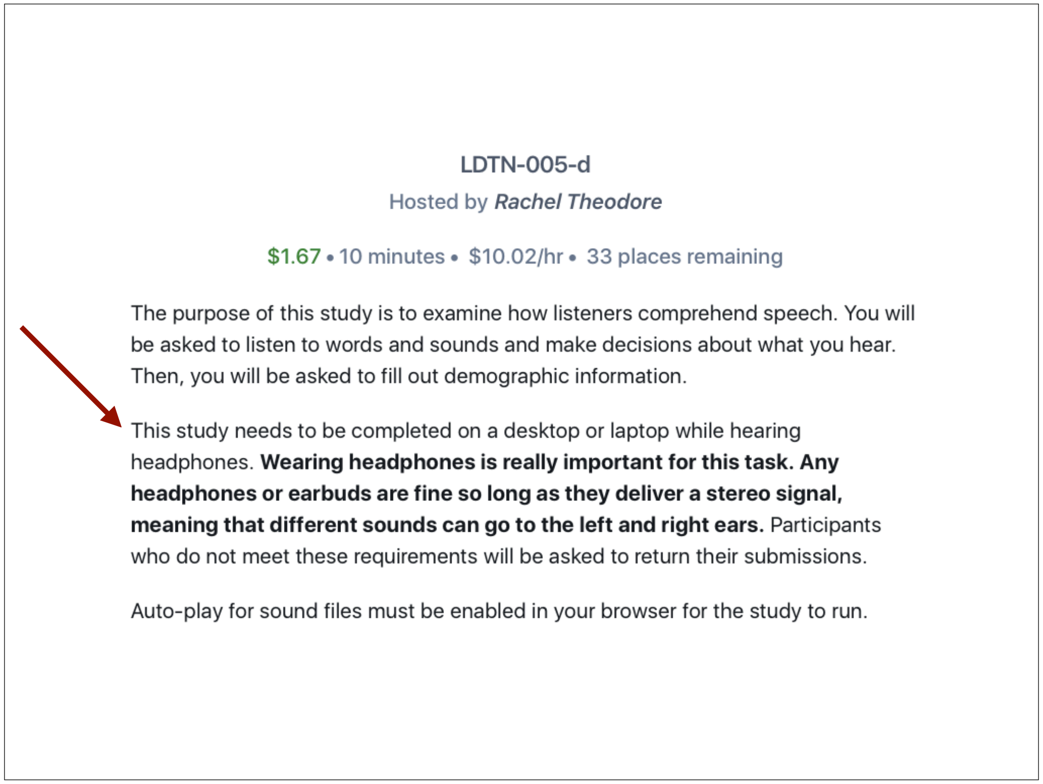
To offset these challenges, here's 12 tips and tricks that we've found useful for promoting high quality data.



Tips and tricks

1. Be *exceptionally clear* with your participants in terms of technology requirements and study instructions

Be very clear with your participants.



LDTN-005-d

Hosted by *Rachel Theodore*

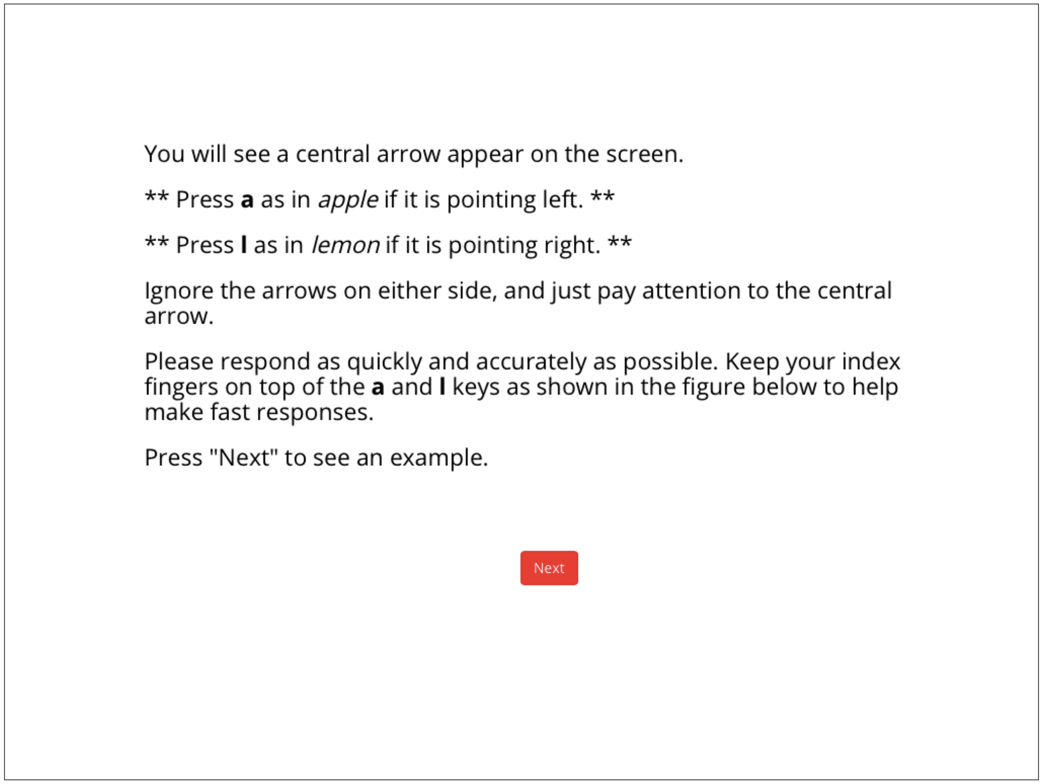
$1.67 • 10 minutes • $10.02/hr • 33 places remaining

The purpose of this study is to examine how listeners comprehend speech. You will be asked to listen to words and sounds and make decisions about what you hear. Then, you will be asked to fill out demographic information.

This study needs to be completed on a desktop or laptop while hearing headphones. **Wearing headphones is really important for this task. Any headphones or earbuds are fine so long as they deliver a stereo signal, meaning that different sounds can go to the left and right ears.** Participants who do not meet these requirements will be asked to return their submissions.

Auto-play for sound files must be enabled in your browser for the study to run.

Let them know what they need to do your study well. Define jargon, like stereo headphones. Tweaking our instructions to provide this definition, really helped to decrease our attrition rate.



You will see a central arrow appear on the screen.

** Press **a** as in *apple* if it is pointing left. **

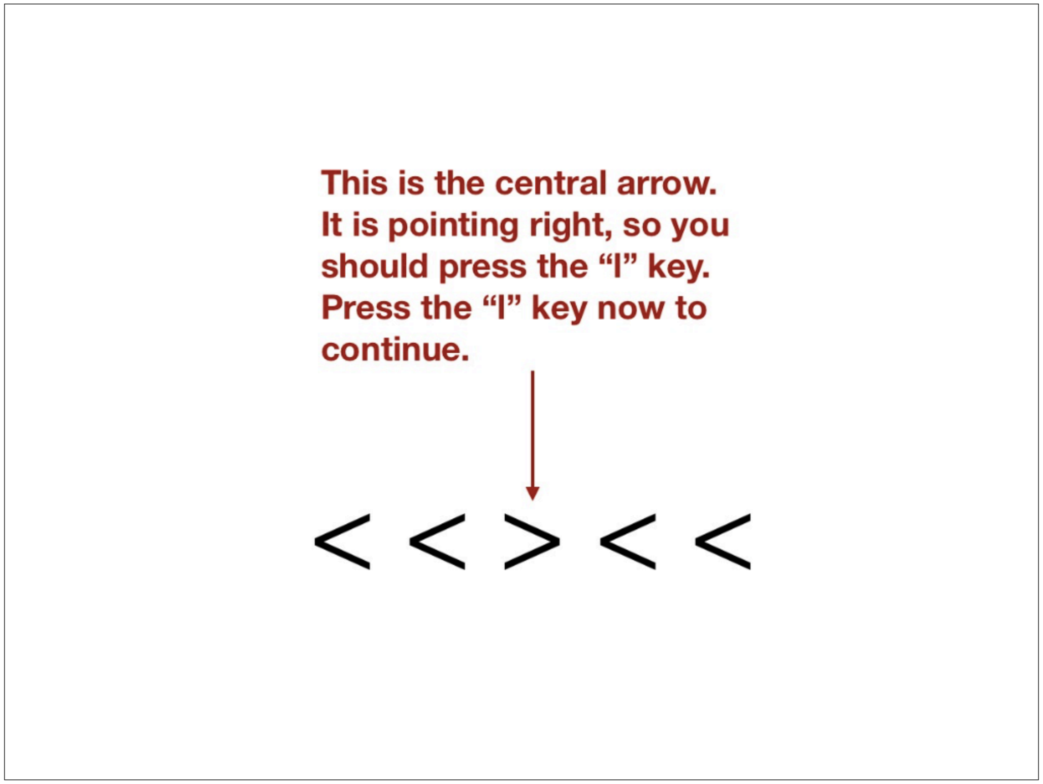
** Press **l** as in *lemon* if it is pointing right. **

Ignore the arrows on either side, and just pay attention to the central arrow.

Please respond as quickly and accurately as possible. Keep your index fingers on top of the **a** and **l** keys as shown in the figure below to help make fast responses.

Press "Next" to see an example.

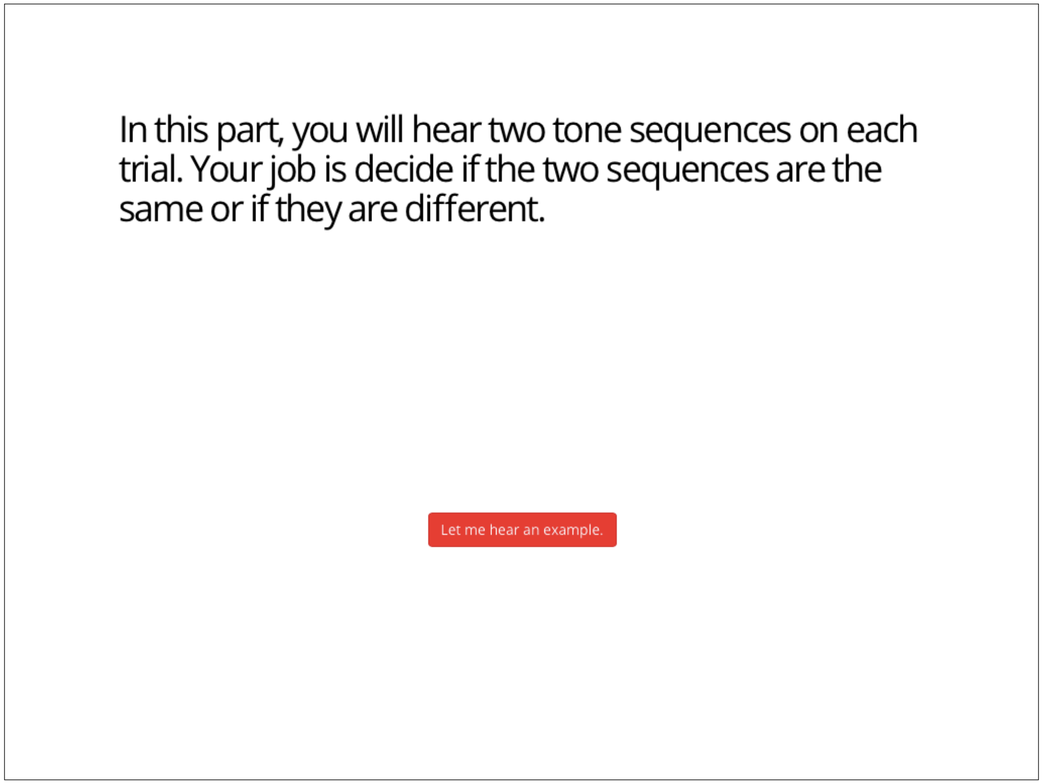
Give guidance; be very clear in your instructions.



This is the central arrow. It is pointing right, so you should press the “I” key. Press the “I” key now to continue.

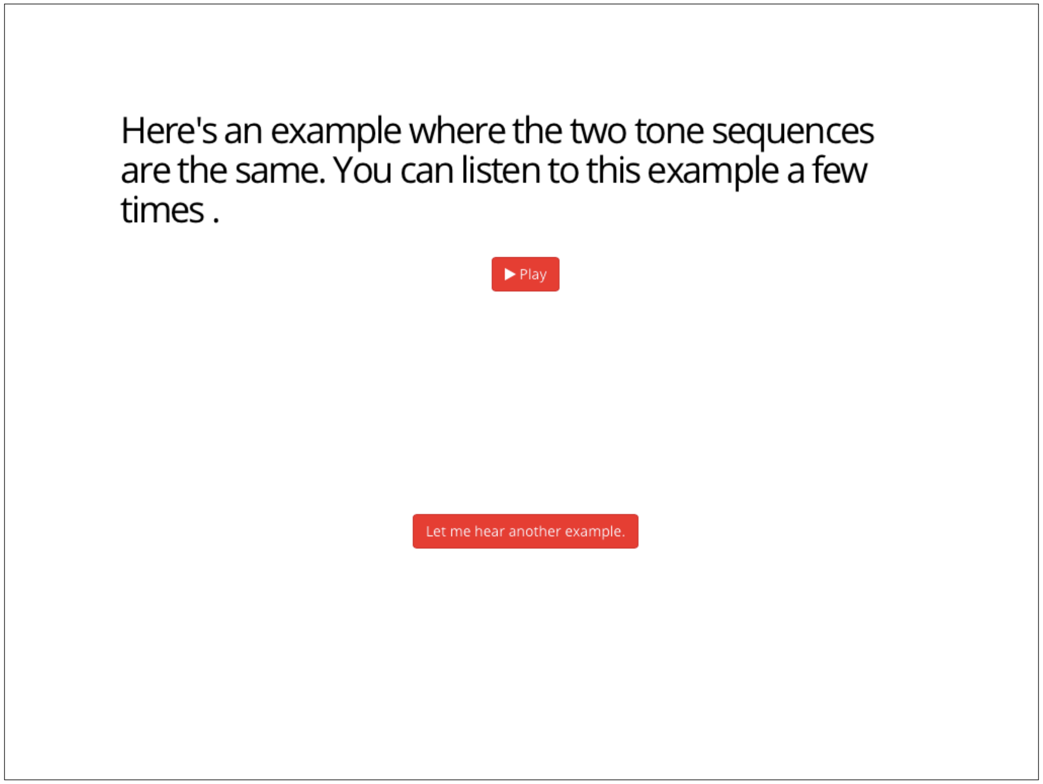
< < > < <

If your task is tricky, let them practice, like we did with the flanker task.

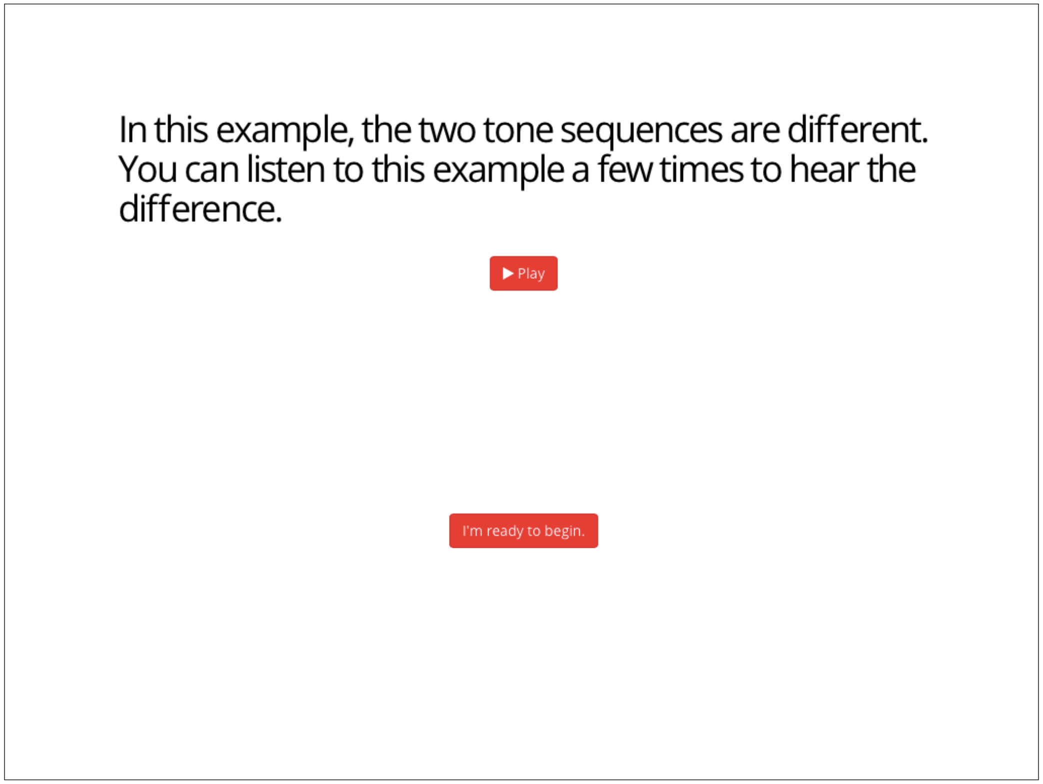


In this part, you will hear two tone sequences on each trial. Your job is decide if the two sequences are the same or if they are different.

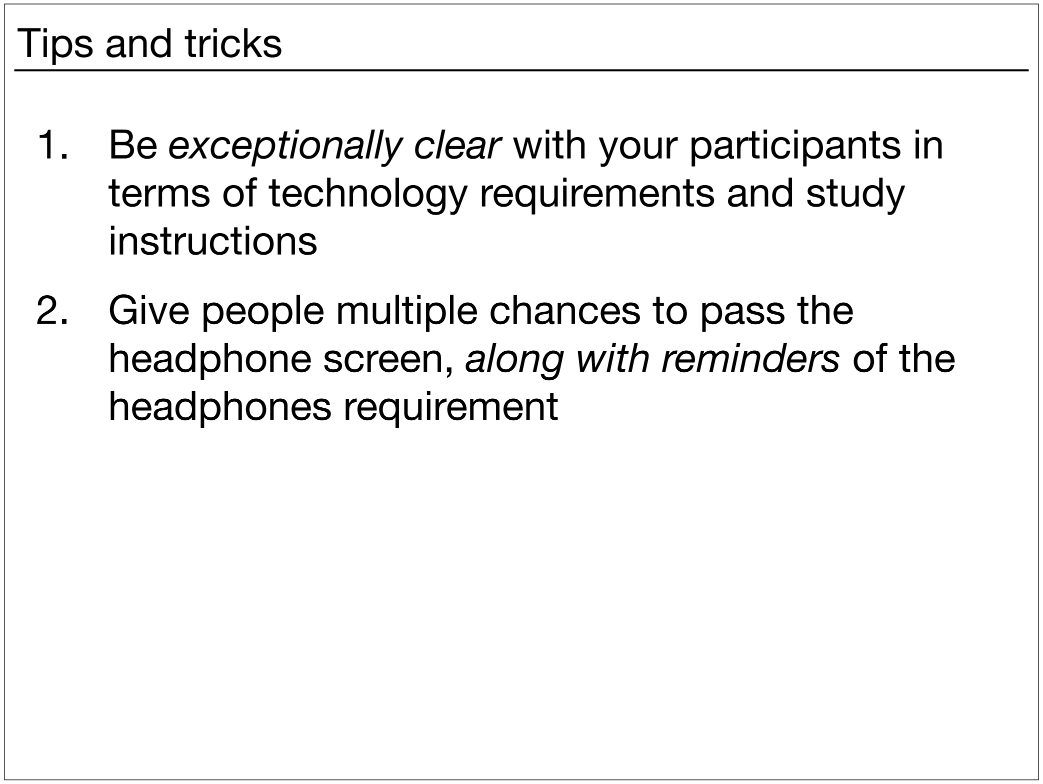
Or like we did for a sound discrimination task.



Here's an example where the two tone sequences are the same. You can listen to this example a few times .



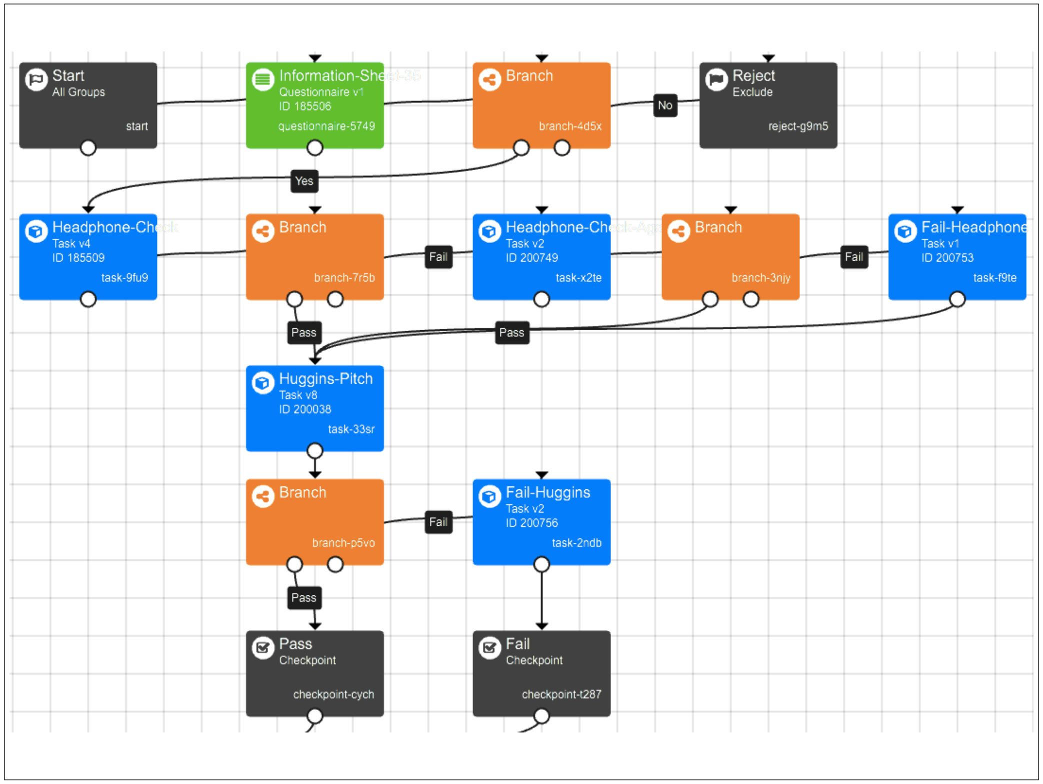
In this example, the two tone sequences are different. You can listen to this example a few times to hear the difference.



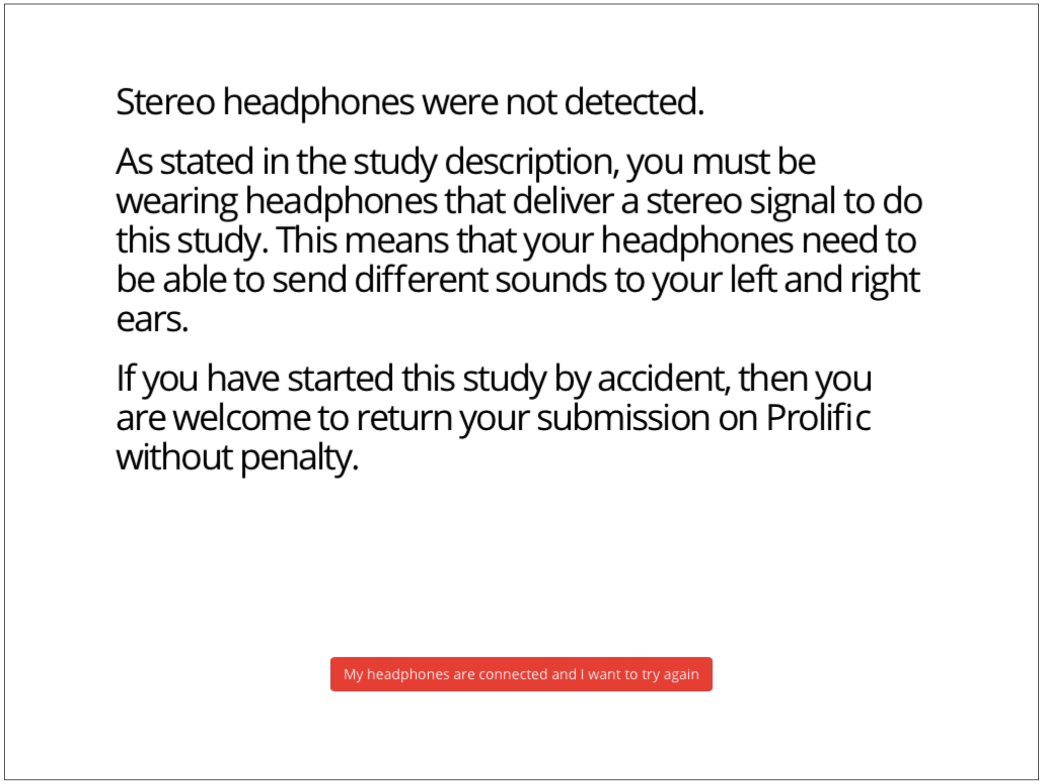
Tips and tricks

1. Be *exceptionally clear* with your participants in terms of technology requirements and study instructions
2. Give people multiple chances to pass the headphone screen, *along with reminders* of the headphones requirement

Give people multiple chances to pass the headphone screen, with a reminder of the study requirements.



It’s easy to set up a branch for this in Gorilla —

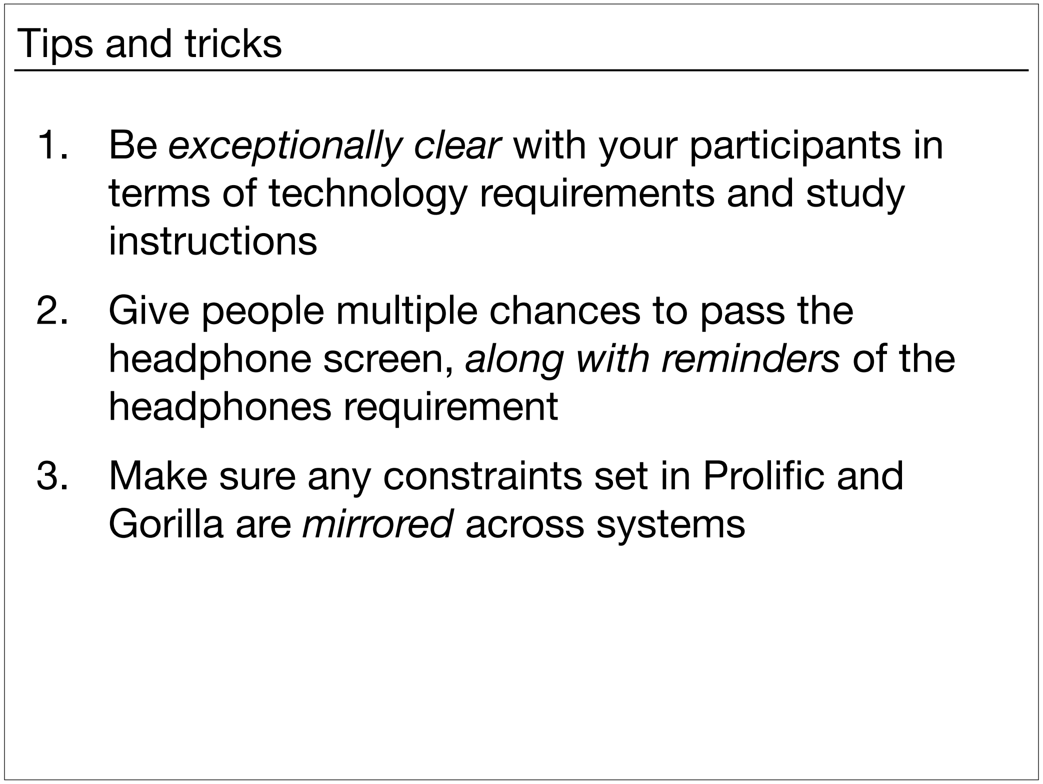


Stereo headphones were not detected.

As stated in the study description, you must be wearing headphones that deliver a stereo signal to do this study. This means that your headphones need to be able to send different sounds to your left and right ears.

If you have started this study by accident, then you are welcome to return your submission on Prolific without penalty.

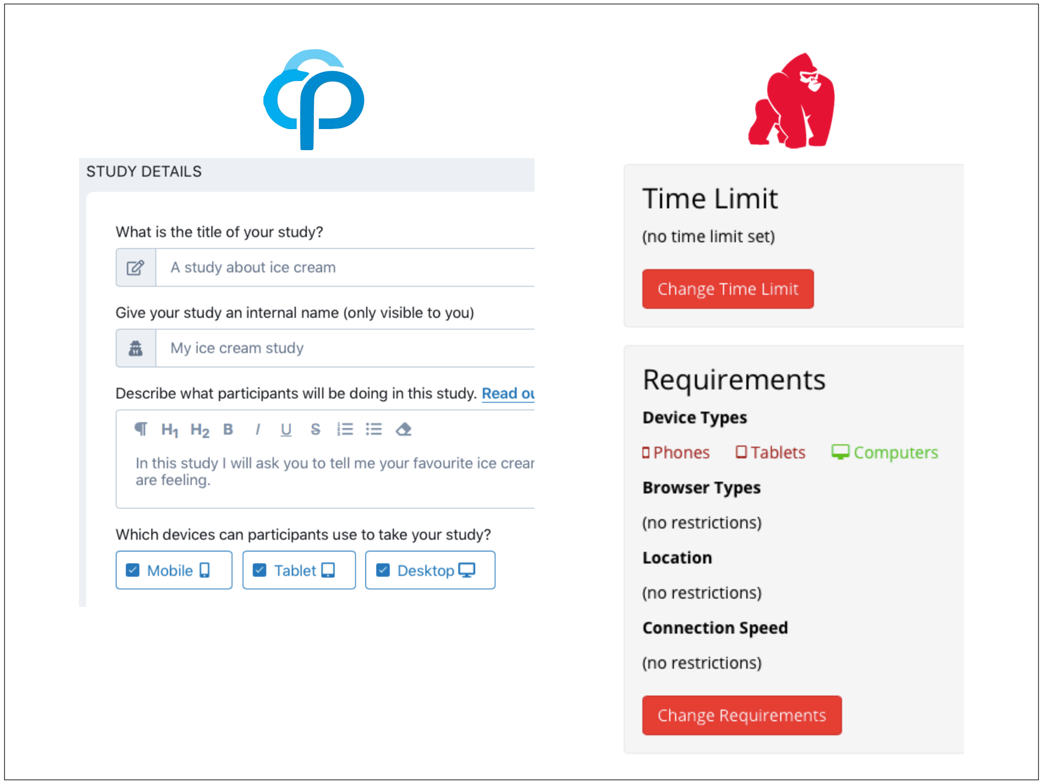
And when we introduced this, our attrition due to lack of headphone compliance went down more than half.



Tips and tricks

1. Be *exceptionally clear* with your participants in terms of technology requirements and study instructions
2. Give people multiple chances to pass the headphone screen, *along with reminders* of the headphones requirement
3. Make sure any constraints set in Prolific and Gorilla are *mirrored* across systems

Be sure to mirror requirements across all systems you’re using.



STUDY DETAILS

What is the title of your study?

A study about ice cream

Give your study an internal name (only visible to you)

My ice cream study

Describe what participants will be doing in this study. Read ou

In this study I will ask you to tell me your favourite ice crear are feeling.

Which devices can participants use to take your study?

☑ Mobile ☑ Tablet ☑ Desktop

Time Limit

(no time limit set)

Requirements

Device Types

Phones Tablets Computers

Browser Types

(no restrictions)

Location

(no restrictions)

Connection Speed

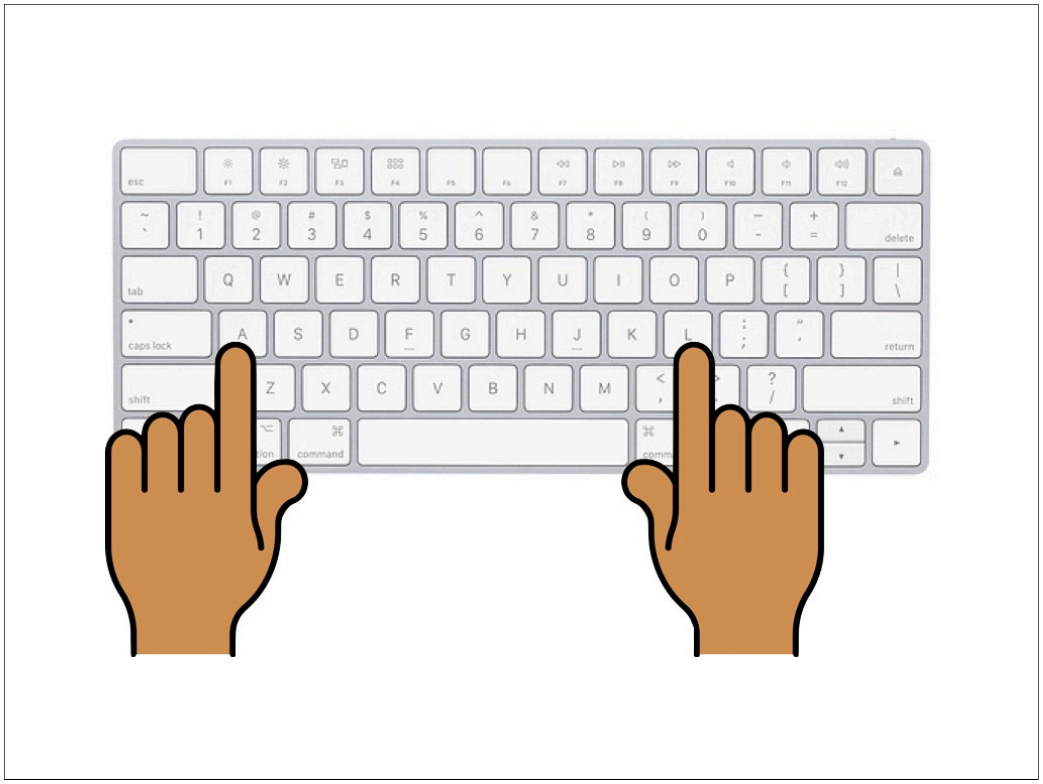
(no restrictions)

For example, if you set a constraint for computer only participation in Gorilla, but don't do that in Prolific too — then Prolific is going to send people to Gorilla only for Gorilla to reject them. This will lead to frustrated participants and a gazillion messages for you to respond to…

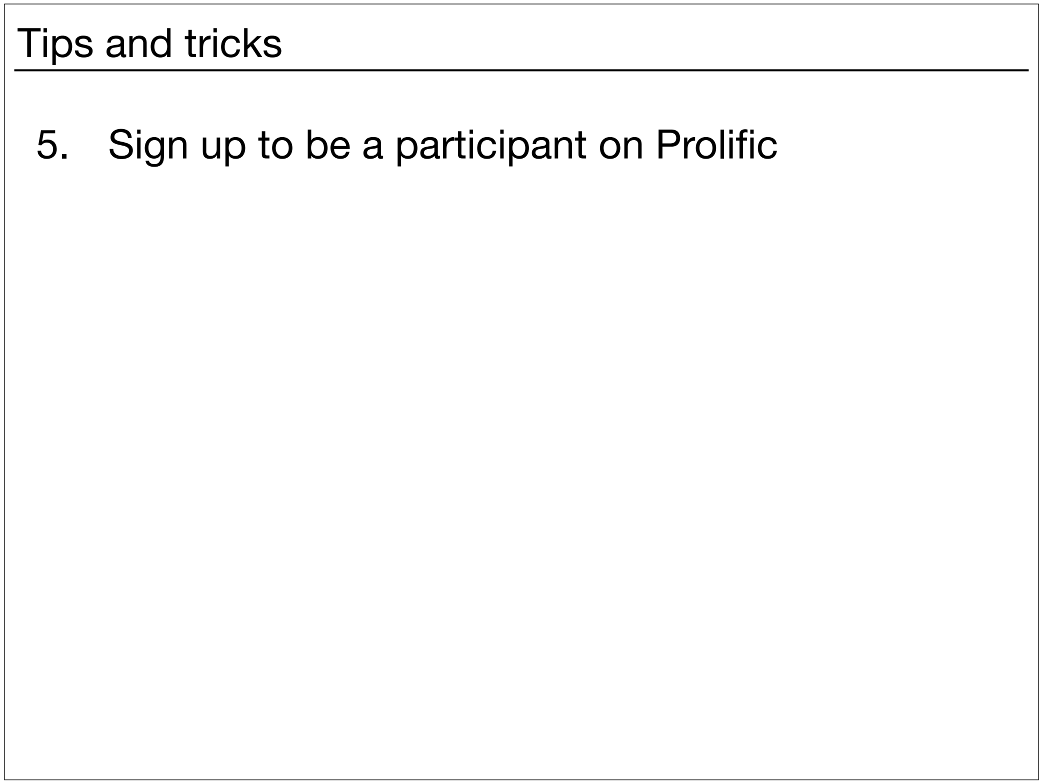
Tips and tricks

1. Be *exceptionally clear* with your participants in terms of technology requirements and study instructions
2. Give people multiple chances to pass the headphone screen, *along with reminders* of the headphones requirement
3. Make sure any constraints set in Prolific and Gorilla are *mirrored* across systems
4. To decrease variability in reaction times, use within-subjects designs and provide a visual cue for hand placement

For reaction time studies, use within-subjects designs when you can, because the largest source of variability is going to come from different hardware/software set-ups across participants.



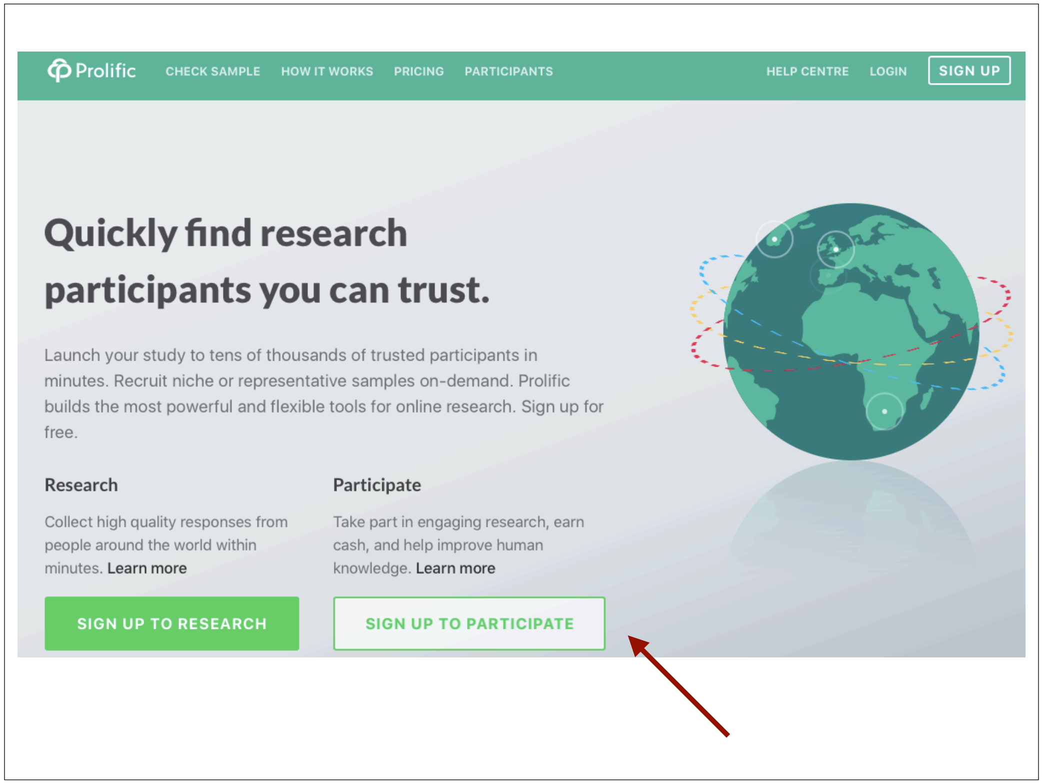
Showing this display for finger placement in RT studies not only led to faster RTs overall, but also less variable RTs.



Tips and tricks

5. Sign up to be a participant on Prolific

Tip 5; join the Prolific pool as a participant yourself!



Quickly find research participants you can trust.

Launch your study to tens of thousands of trusted participants in minutes. Recruit niche or representative samples on-demand. Prolific builds the most powerful and flexible tools for online research. Sign up for free.

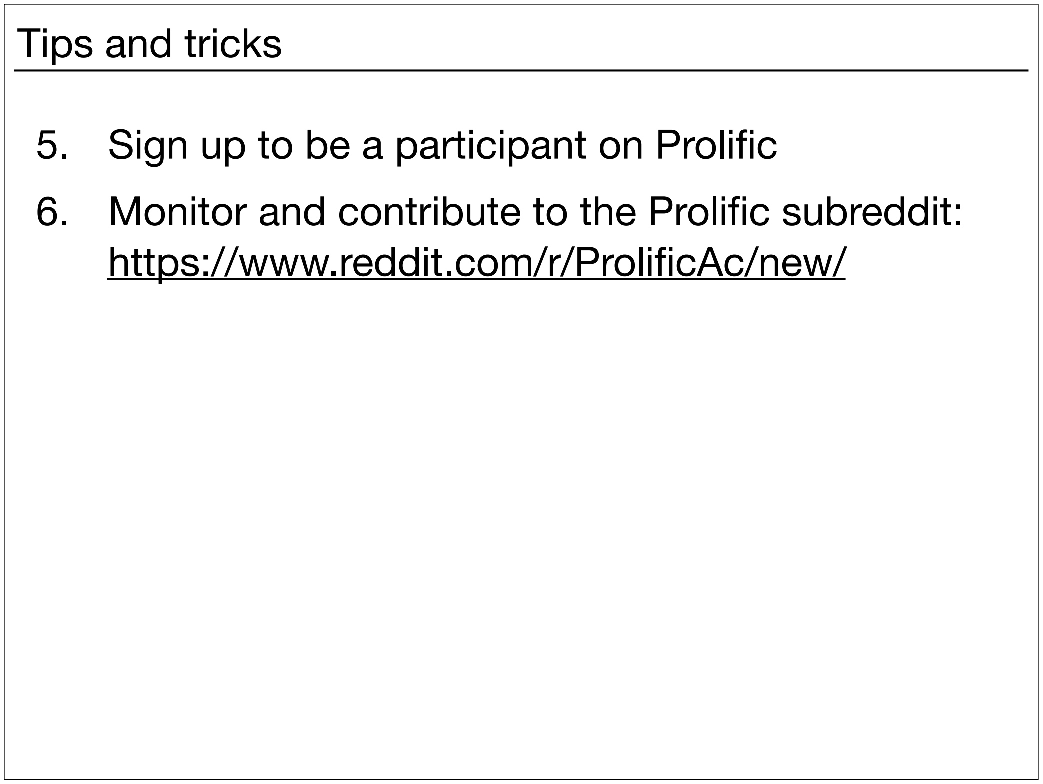
Research

Collect high quality responses from people around the world within minutes. **Learn more**

Participate

Take part in engaging research, earn cash, and help improve human knowledge. **Learn more**

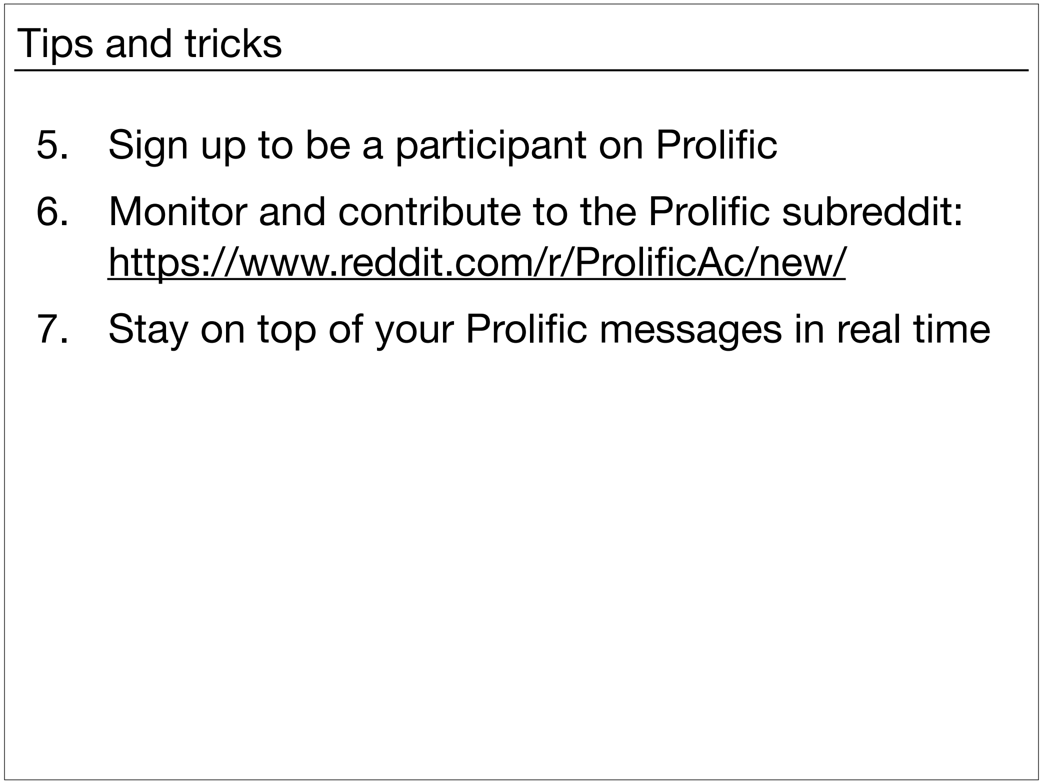
I’ve learned so much about how to make better studies that lead to a better participant experience by being a participant myself.



Tips and tricks

5. Sign up to be a participant on Prolific
6. Monitor and contribute to the Prolific subreddit: https://www.reddit.com/r/ProlificAc/new/

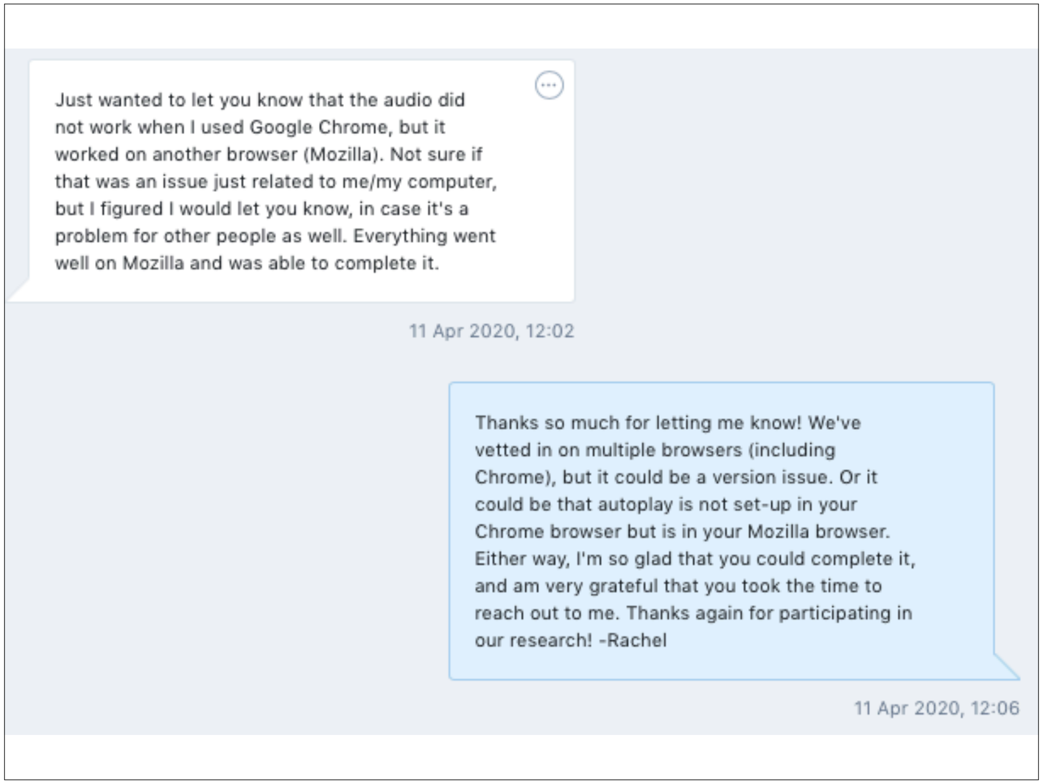
Related, monitor and contribute to the Prolific subreddit; I can’t stress enough how much we can learn from our participants! You’ll quickly tune in to what drives them crazy, what they enjoy — and then you can use this information to optimize your designs.



Tips and tricks

5. Sign up to be a participant on Prolific
6. Monitor and contribute to the Prolific subreddit: https://www.reddit.com/r/ProlificAc/new/
7. Stay on top of your Prolific messages in real time

Prolific has a great messaging system for researchers and participants. Stay on top of those messages!



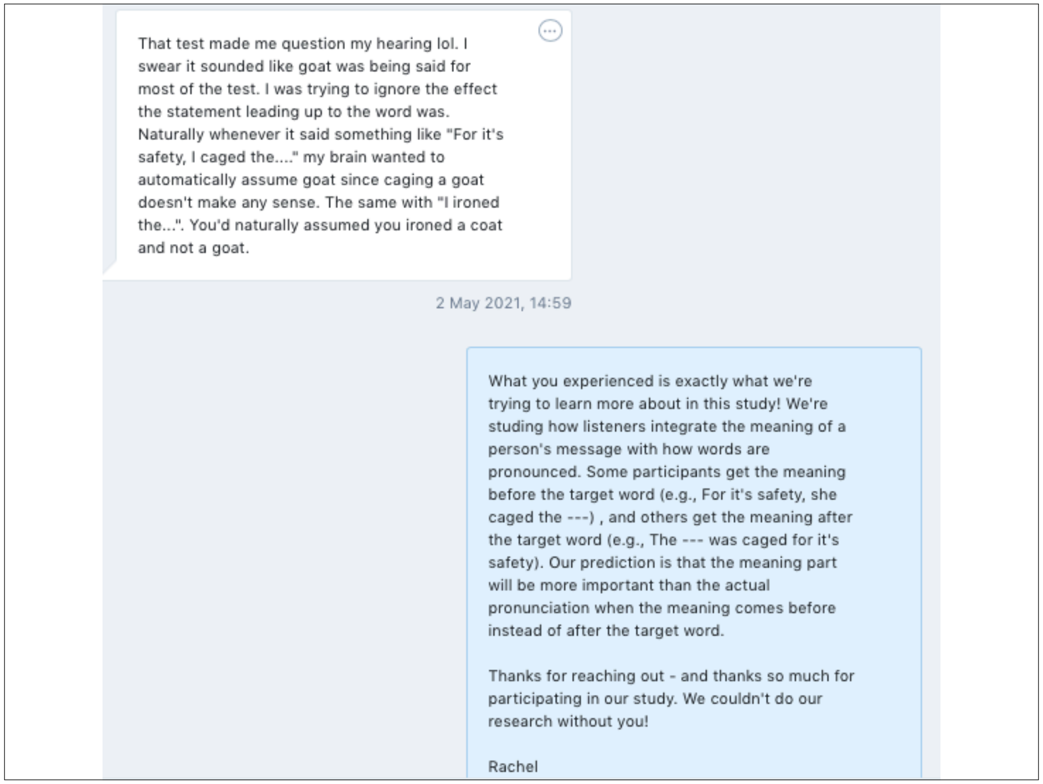
Just wanted to let you know that the audio did not work when I used Google Chrome, but it worked on another browser (Mozilla). Not sure if that was an issue just related to me/my computer, but I figured I would let you know, in case it's a problem for other people as well. Everything went well on Mozilla and was able to complete it.

11 Apr 2020, 12:02

Thanks so much for letting me know! We've vetted in on multiple browsers (including Chrome), but it could be a version issue. Or it could be that autoplay is not set-up in your Chrome browser but is in your Mozilla browser. Either way, I'm so glad that you could complete it, and am very grateful that you took the time to reach out to me. Thanks again for participating in our research! -Rachel

11 Apr 2020, 12:06

In my experience, participants are quick to report when something goes wrong.



That test made me question my hearing lol. I swear it sounded like goat was being said for most of the test. I was trying to ignore the effect the statement leading up to the word was. Naturally whenever it said something like "For it's safety, I caged the...." my brain wanted to automatically assume goat since caging a goat doesn't make any sense. The same with "I ironed the...". You'd naturally assumed you ironed a coat and not a goat.

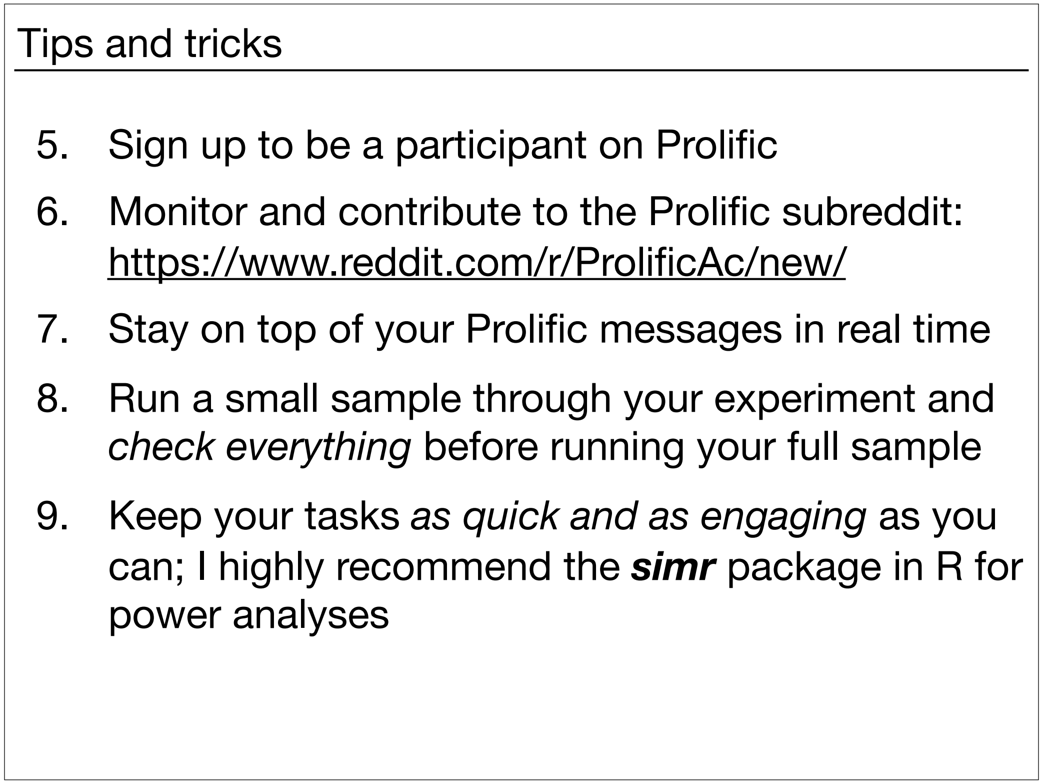
2 May 2021, 14:59

What you experienced is exactly what we're trying to learn more about in this study! We're studing how listeners integrate the meaning of a person's message with how words are pronounced. Some participants get the meaning before the target word (e.g., For it's safety, she caged the ---) , and others get the meaning after the target word (e.g., The --- was caged for it's safety). Our prediction is that the meaning part will be more important than the actual pronunciation when the meaning comes before instead of after the target word.

Thanks for reaching out - and thanks so much for participating in our study. We couldn't do our research without you!

Rachel

AND — messaging with your participants is an excellent forum for science communication.

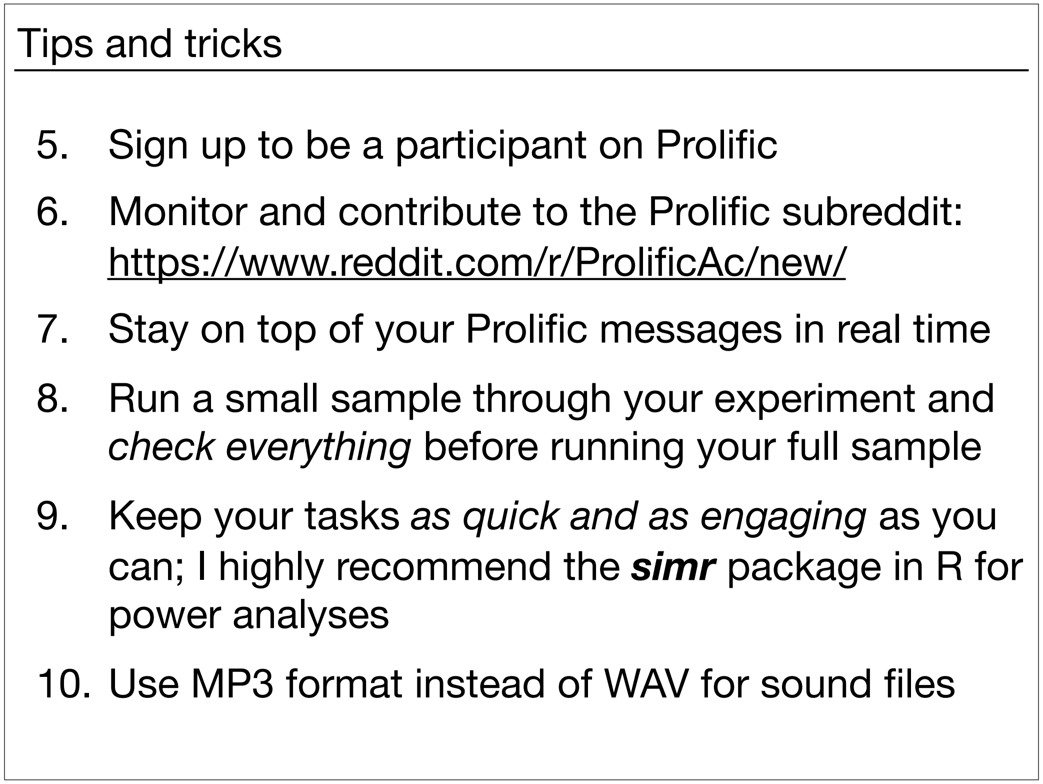


Tips and tricks

5. Sign up to be a participant on Prolific
6. Monitor and contribute to the Prolific subreddit: https://www.reddit.com/r/ProlificAc/new/
7. Stay on top of your Prolific messages in real time
8. Run a small sample through your experiment and *check everything* before running your full sample
9. Keep your tasks *as quick and as engaging* as you can; I highly recommend the ***simr*** package in R for power analyses

A few more tips and tricks — run a small sample through before you run the full sample. The only downside of being able to collect data from 300 participants in an hour is that one is also able to make a fatal mistake that affects 300 participants in an hour…

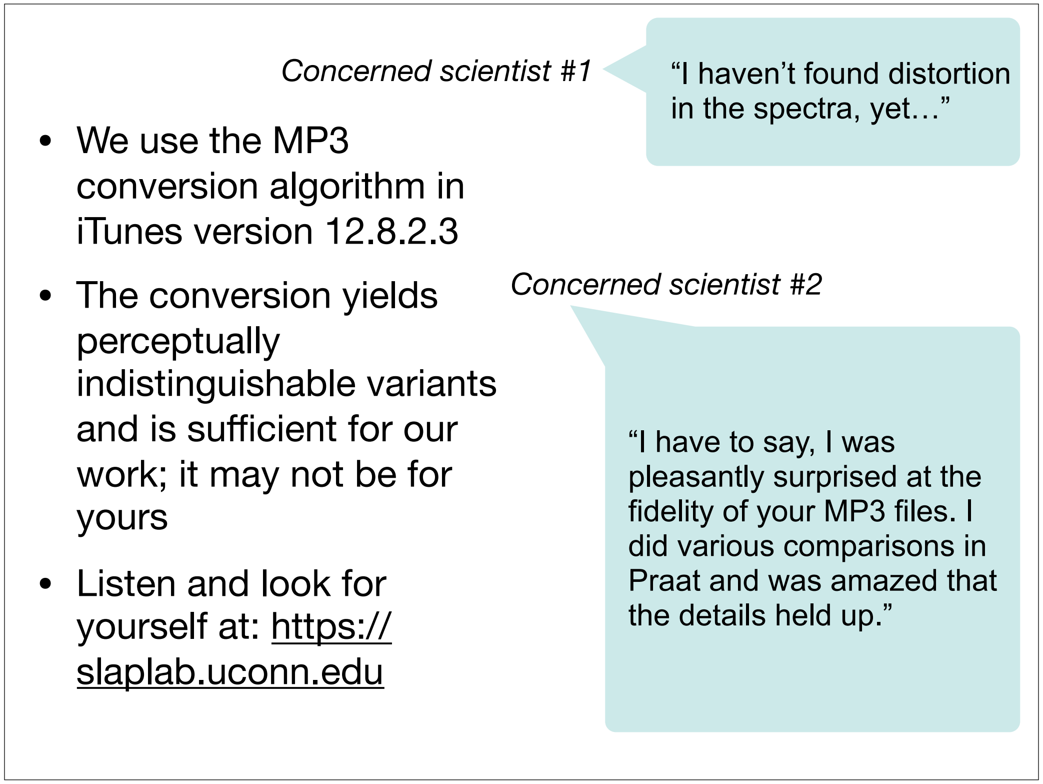
Keep your tasks as quick as you can while also ensuring adequate power. Gorilla has just released a game builder feature that I’m really excited about as a means to make our boring psychophysical tasks more engaging for participants, which will only benefit data quality.



Tips and tricks

5. Sign up to be a participant on Prolific
6. Monitor and contribute to the Prolific subreddit: https://www.reddit.com/r/ProlificAc/new/
7. Stay on top of your Prolific messages in real time
8. Run a small sample through your experiment and *check everything* before running your full sample
9. Keep your tasks *as quick and as engaging* as you can; I highly recommend the ***simr*** package in R for power analyses
10. Use MP3 format instead of WAV for sound files

Tip 10; use MP3 format instead of WAV for sound files.



Concerned scientist #1

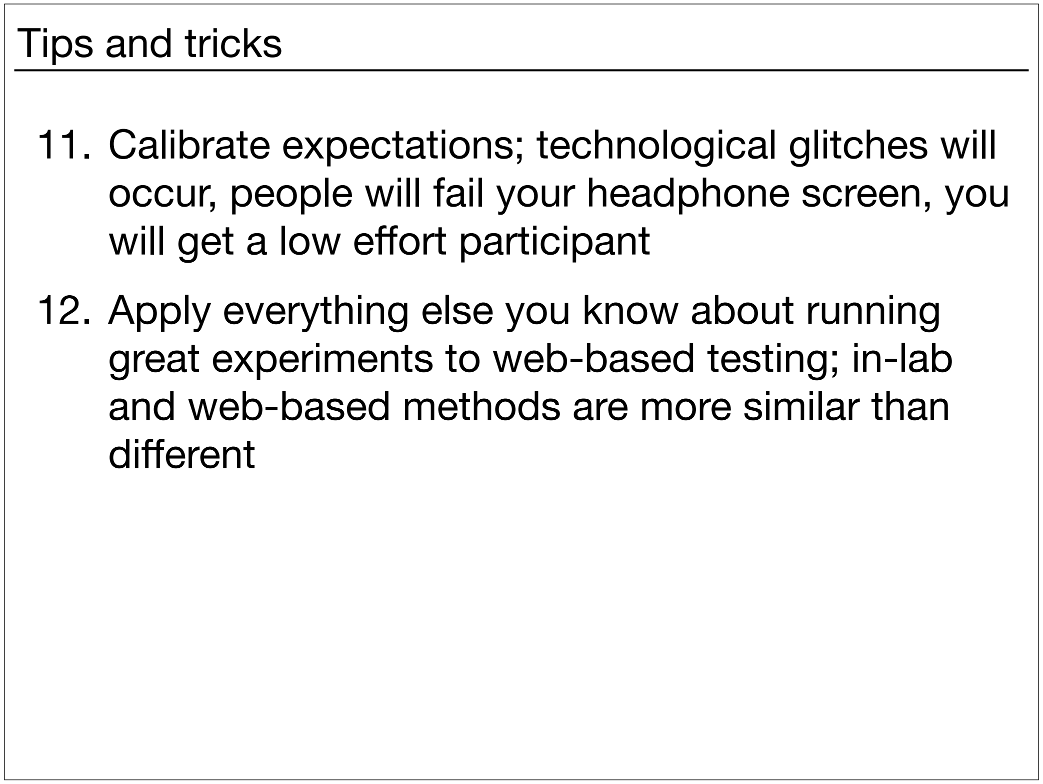
"I haven't found distortion in the spectra, yet…"

- We use the MP3 conversion algorithm in iTunes version 12.8.2.3
- The conversion yields perceptually indistinguishable variants and is sufficient for our work; it may not be for yours
- Listen and look for yourself at: https://slaplab.uconn.edu

Concerned scientist #2

"I have to say, I was pleasantly surprised at the fidelity of your MP3 files. I did various comparisons in Praat and was amazed that the details held up."

MP3 is native to browsers and you will run into glitches with some participants if you use WAV files. I know, I know, we've all been trained to avoid lossy formats. BUT, conversion algorithms are very good these days and I've yet to find anyone who can detect important missing information in our MP3 conversions. You can try yourself with the examples on our website. All of the studies I showed in this presentation used MP3 audio stimuli.

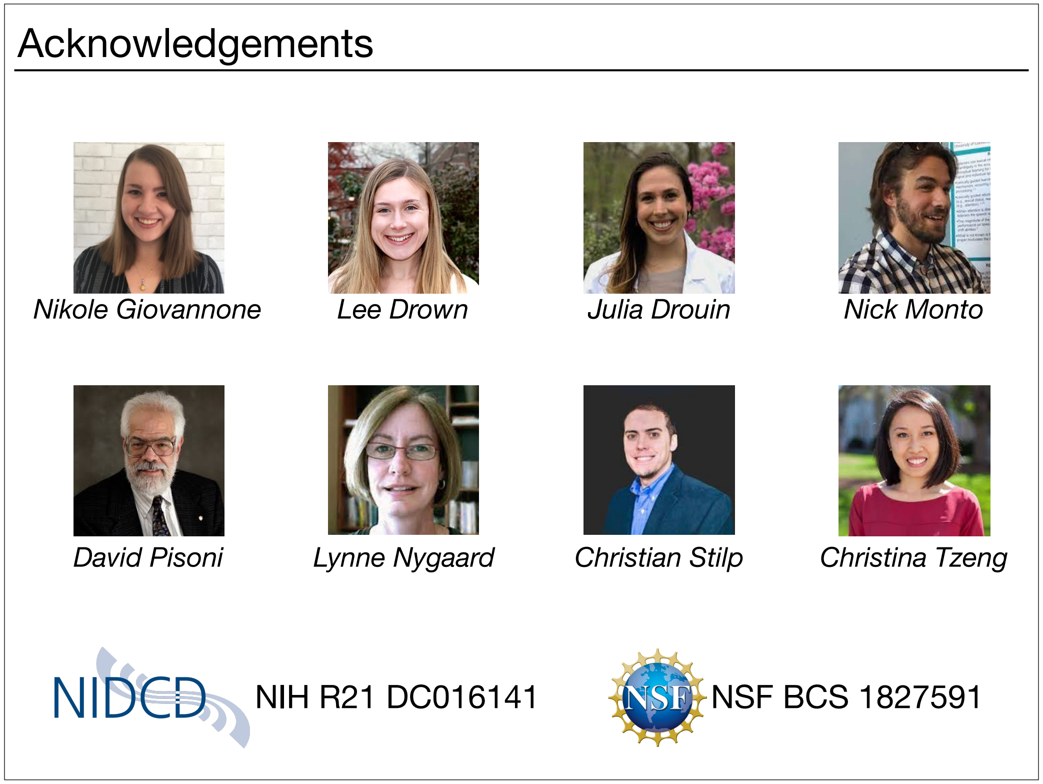


Tips and tricks

11. Calibrate expectations; technological glitches will occur, people will fail your headphone screen, you will get a low effort participant
12. Apply everything else you know about running great experiments to web-based testing; in-lab and web-based methods are more similar than different

Last two tips: Calibrate your expectations; you’re going to have glitches, you’re going to have a low effort participant; these things happen even in the laboratory. Look for them; and design tasks that make it easy to detect low effort responses.

And finally — don’t forget to apply everything else you know about running great experiments; in-lab and web-based methods are more similar than different.



Acknowledgements

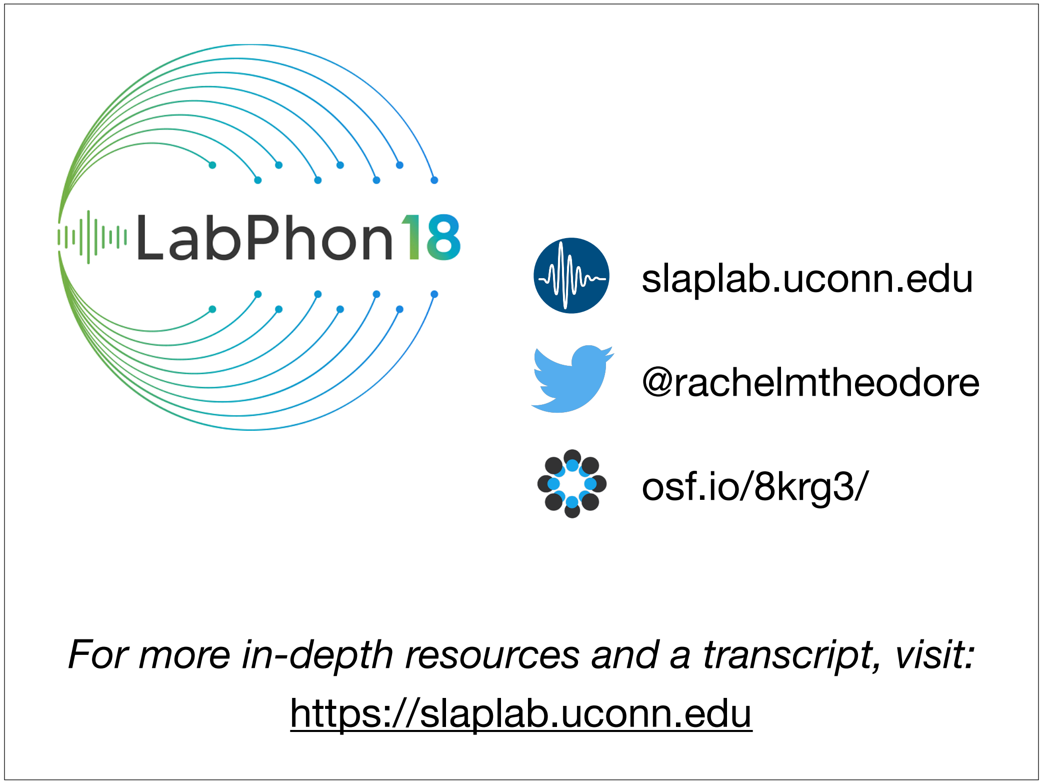
Nikole Giovannone *Lee Drown* *Julia Drouin* *Nick Monto*

David Pisoni *Lynne Nygaard* *Christian Stilp* *Christina Tzeng*

NIDCD NIH R21 DC016141

NSF NSF BCS 1827591

With extreme gratitude, I acknowledge the trainees and colleagues who have been instrumental in making our foray into web-based methods a success, and the funding sources for our work.



LabPhon18

slaplab.uconn.edu

@rachelmtheodore

osf.io/8krg3/

For more in-depth resources and a transcript, visit:
https://slaplab.uconn.edu

Additional resources are available on our website and OSF page; and please don’t hesitate to reach out to me offline if you have questions. Thank you very much.