Protocol for Remote Testing for the Second Language Psycholinguistics Lab

Posted version date June 9, 2020

Second Language Psycholinguistics Lab, Indiana University

Contents

The present document is intended as a guide for conducting behavioral experiments online through a web browser, in the field of language processing, speech perception and production, involving precise stimulus presentation controls and response (error or reaction time [RT]) collection, as well as the need to control some environmental factors during the experiment.

Remote testing, online testing: definitions

Remote testing refers to conducting behavioral experiments online through a web browser, without the researcher and participant being physically present in the same space.

Online testing refers to conducting behavioral experiments online through a browser. This can happen remotely as above, but it can also happen in person, face-to-face, using a web browser for stimuli presentation and data collection.

Remote testing: Why?

Given the uncertainty due to the current Covid-19 outbreak (status: April, 2020), we may not be able to do in-person experiment for a while. It is important to equip ourselves with techniques and principles for conducting real-time experiments online through remote testing, as a way to continue collecting data for our research.

Yet there are other reasons as well. Beyond this particular situation, doing experiments online can open up access to populations that we previously considered "unrealistic," for example because it would involve travelling or because their numbers were few on our campus.

Finally, being able to conduct experiments online (not remotely), that is in person in the lab or on campus, but *through an online platform,* can also alleviate some of the issues we've ran into with different computers across labs on campus and the availability of different software versions for example. Using a browser and testing online does not remove this problem entirely, and it also brings about other specific problems we need to solve, but in some cases, it can help. Having the choice is usually the better option.

Remote testing: How?

This question is the main reason for this document. How do we, as a lab, develop a coherent approach to online testing? What safeguards to we need to put in place to guarantee the highest quality data, the reproducibility of data, and ultimately, trust that our results are worth our effort?

The purpose of this document is to establish a protocol that we can all follow, that will ensure the highest quality data obtained through remote testing.

Web browser experiments are not new. However, there are very few speech perception experiments being done, and very few experiments with a specific target group such as second language (L2) learners of a certain language. The main reason behind it is that recruiting a very specific group of people and obtaining reliable data involving speech sounds remotely is very difficult: the mass-recruitment of participants through platforms such as **Sona, Prolific, or Mechanical Turk** involves very low control over who is actually participating, and placing too strict restrictions or having a complex series of tasks is unrealistic using these platforms.

For our lab, we need a different, more tailored approach. Obviously, we have a lot less control over some important questions (such as what equipment is being used, who is the participant, what are the conditions in which the experiment is run, how to motivate participants to stay on task, how to run several experiments in a row) in a remote testing situation compared to an in-person testing in the lab, individually, with the lab equipment that is held constant. There will be some sacrifices that we will need to make. There will probably also be more "participant loss" due to people not completing the experiment, to technical glitches, wifi crashes, etc. We need to be prepared for it, but we can mitigate the problems. This document will help us decide what is possible.

Right now, all over the web, experiments abound. Some of them are very hands-off (researchers never contact the participant, it's fully anonymous because essentially random people click on a link, they do not know who does the experiment) to very hands-on (not anonymous recruitment, individual zoom meeting where the answers are gathered in the presence of the researcher).

For our typical speech perception experiments, we decided to opt for a **middle-ground** between these two extremes, which is in fact closer to the "hands-on" end of the continuum. We use **targeted recruitment** via professional contacts or field-specific distribution lists, as well as "word of mouth" and recruitment through social media and acquaintances, but there is a gateway: interested participants are contacted with more information on the study, and an appointment is made, during which the experiments are explained and consent is obtained. Each participant is then individually guided through the experiments, strictly following the approved experimental protocol, up to the debriefing and – if applicable – payment via e-gift cards. Targeted questions are used to evaluate the conditions in which someone is doing the experiment, and a background questionnaire is used to evaluate the characteristics of the participant.

The rest of this document describes some of the procedures we have agreed upon for our lab.

List of recommendations for remote testing studies

Generalities – important aspects to implement for all new studies

Given that a study done remotely takes more time and requires more questions and control tasks, the following points are important to keep in mind

- **Create studies that are <u>less</u> ambitious** in scope generally. Adopt a more cumulative approach, smaller studies but that build on one another.
- **Aim do design "compact" studies**: tasks should take as little time as possible, the set up should be as simple as possible (as few tasks as feasible). This helps with keeping people motivated and concentrated on the task.
- **Be careful of file size** – files should be as small as possible, in order to reduce the time it takes to download an experiment for each participant
- **Adopt clear and consistent file naming** conventions with version tracking

- examples: *2020may18_litreview.docx* or *litreview[v3].docx* ; *2020apr23_ oddityana.xlsx; 2020feb22_minutes.docx*
- For lab specific guidance, see internal "file naming conventions for SLPL" document

Alterations to the experiments and testing themselves

- **Create a virtual meeting room for each person (using Zoom, Skype or another app)**, and stay there for the duration. If you have several tasks, the person can leave and rejoin the meeting, or can just turn off sound and video and leave the meeting "on". You turn off your sound and video as well. Also provide your cell or another method for contacting you in case their internet/computer crashes. Send them the info in an e-mail separately.
  - We have to do this because the fully "hands-off" testing will not work for our research. We rely a lot more than other studies on population controls, that is, who is taking the experiment. Hence our long and complex background questionnaires for instance.
  - Goal: Striking a good balance between hands-on and hands off. We must connect with people and test them individually, which means exactly that: we meet with them, walk them through, and leave them alone for the main experiment, with a debrief at the end.
    - Added bonus: this acts as a great motivator to help people complete the tasks
    - Require participants to put their phone out of sight, turn it off or on silent, and to be in a place where they can focus (few distractions).
- **Create a written protocol document and follow it to the letter while you test, in order to maintain the same experience for everyone as much as possible**
  - Practice all instructions with pilot subjects, to verify that they can do it on their own, or that your instructions are sufficient, and practice potential bugs or issues
  - This step requires some serious debugging several times until an optimal sequence of instructions is found – a bit like trying out a cooking recipe with various people. Do this, with several people at first, to fine-tune the exact protocol that will work for most people
  - Try to anticipate road blocks and issues, and have a plan B to work around them
- **Keep a log ("testing log")** of what happens every time you test, writing down especially if the situation departs from the protocol in any way
- **Implement requests for optimizing browser** and computer performance
  - Internet stability issues: Psychopy tests "locally" because it uploads all the files first, and then runs. So the internet stability during the experiment itself is not such a huge deal. But it's good to ask people which browser they use (see above).
  - Make recommendations at the beginning of the zoom session that people close down all apps, like dropbox, close any other applications or music during the experiment etc., also remove multiple tabs on browser, only have one browser open, and close all applications that they don't need.
    - This may also be a reason to actually leave the zoom meeting, instead of having it run in parallel, at least for perception experiments. We are still experimenting to what extent having the virtual meeting room running slows down the tasks.

- **Add questions for the** "**Distraction check**" either at the debriefing meeting, or as a block of questions in Qualtrics [we have created these for sharing: "Remote Testing Questions"; See pdf]
    - Normally, we did this through the testing log, where we would jot down any issues that occured during the experiment. Since we're not in the physically same location in remote testing, this is not possible, so we now ask the participant to give us information about this afterwards.
    - If you pay the participants, make sure they understand that the payment does not depend on the answers.
    - The questions ask about the following
        - Rate the environmental noise (mostly quiet ----- pretty loud all the time); for perception and production studies, we use a determined threshold for inclusion of dataset based on the answers.
        - Rate their level of concentration during the task (not at all, very concentrated ---------- quite distracted, not focused)
        - Did you find the experiment difficult? evaluate the level (pretty easy ---- pretty difficult)
        - Did anything happen during the experiment? (interruption, loud noise, phone rang and I answered, kept getting texts…)
- **Add questions for the** "**context/environment check**" [Now a block of questions in Qualtrics "Remote Testing Questions"; See pdf]
    - 1) I'm currently : where the person is (choices: at home /office-study/ studio/public place/ on the go)
    - 2) I'm using a : how they're typing (choices: smartphone / tablet / laptop / game console / desktop computer)
    - 3) I'm hearing sounds through: how they're getting audio input (choices: ear buds / headphones / built-in loudspeaker / TV / external loudspeaker). *=> only needed for studies using sound.*
    - It is also important to
        - Check some details about the type of input device (e.g. Do you have a USB keyboard? a touch screen?)
        - Verify that the crucial keyboard buttons (e.g., the space bar) are working well on the person's keyboard
- **Add questions for the "Operating system check"** [Metaquestion in Qualtrics if using]
    - Qualtrics has a "metaquestion" (https://www.qualtrics.com/support/survey-platform/survey-module/editing-questions/question-types-guide/advanced/meta-info-question/) which logs which browser, version, OS, screen resolution were used during the experiment. The only assumption is that they use the same browser to answer the Qualtrics questionnaire and the other task such as the perception experiment. This can be required or the two tasks (the questionnaire and the task) can be connected internally through a redirect in Qualtrics.
    - If not using Qualtrics:

        - Which computer OS are you using? Forced Choice: Windows 7, Windows 10, Mac, Ubuntu… (add others for smartphone and tablets)
        - Which browser are you using? (depending on their answer above, dropdown will be specific to windows or Mac, for instance; Firefox, Safari, Chrome, Edge…)
        - It may be important to also consider more "obscure" browsers or to add an "other" option with text input.
        - It also may be important to add an option "I don't know".
    - These may seem not very useful now because we still do not know how to interpret or what to make of the information, but they may turn out important if some research shows later that this matters. We'll then have the information to look into it.
- **Add the "Sound check"**
    - *Extra very short task (2-3 minutes) that requires people to press the space bar to detect a tone.*
    - *We are currently working on a Psychopy version of this task.*
- Use more **breaks**
    - *good to avoid distractions from interfering with experiment*
- **Consent is provided at the beginning of the study** through a screen showing the SIS (see example), or – if not possible, send the SIS up front as a "Experiment sheet" for the person to read while you are with them through the zoom meeting. Obtain verbal consent with an explicit question.
- **Implement a final debrief**, where you also settle any payment option such as e-gift-card number/payment (see below, section "Payment")
- Optional: To increase motivation during the task, if you turn off your video, you can also display a motivational image of you or something like that.

Participant payment

In many of our studies, we pay participants for their time. This is usually about $10/hr. If recruiting in classes, we are also sometimes able to arrange for extra credit. Currently, with the move to online of all teaching, the option of recruiting in classes is temporarily more complicated (not impossible, but worth the hassle?). This can change soon though – and extra-credit can again become a good option.

For payment, we use electronic gift cards: the cards are purchased in advance through the department and the participant is given their code to use with a specific website or provider (e.g. Amazon, Kroger, or other e-gift-cards that work for the participant's given environment).

Technology and logistics

Our lab uses the **Pavlovia interface** (Pavlovia.org) and compatible experimental software such as Psychopy or jspsych and Qualtrics for surveys. The participant can use a variety of browsers although we recommend one of the most common ones (Google Chrome, Firefox, or Safari). Other experimental

formats will rely on other tools as needed (Praat, for instance) or other existing third-party online tasks (for example vocabulary knowledge tasks).

Using a (shared) Pavlovia account

If using the Pavlovia interface to test remotely, know that for each participant who completes the study, Pavlovia will charge a participant "credit". The standard charge for use of Pavlovia is through Participant Credits, which have to be bought in advance from the Pavlovia Store. https://pavlovia.org/docs/store/pricing . Credits can be purchased on the website and currently cost about 25 cents (in US$). So, for 60 participants, this would amount to $15.

Things to keep in mind when using a shared Pavlovia account (for the lab, for example):

- The experiment name must be the same as the folder that it is based on (on your computer).
- Make sure someone else has a full backup of the complete experiment in its FINAL, what we call "posted", version. Never modify this posted version after the experiment has started.
- When creating lots of experiments in one account (the lab's), it's CRUCIAL to have clear names for them.
- **Naming guidelines**: In our lab, we use the Year/Month, Lab member initials who uploaded/created it, and short experiment name, **for example:**
    - **20-05ID_LexDecPortuguese** (for a lexical decision task on Portuguese, created in May 2020 by Isabelle Darcy)
    - **20-07ID_MenLexOrthography** (for a series of 2 experiments on the role of orthography for the L2 Portuguese mental lexicon, created by Isabelle Darcy in July 2020)
- Since we will ALL use the same account (SLPL, under my name) for Pavlovia, all updates and "new commits" (edits) will all be "named" with my name ("Isabelle Darcy updated file 2 on April 18" for example). For this log info to be useful, we ALL need to **initial each change we make** in the comments.
- We also have to be very careful not to edit, download or modify anything that's not "ours" since with the lab credentials, anyone can edit, download, delete, and take over the admin rights of any experiment. This is a bit scary at this point, so we are looking into perhaps making folders or groups. Ultimately, perhaps the one common account is not the ideal solution (unless we're all very disciplined), but all of us having individual accounts also has drawbacks and it means more admin no matter what.

Recommendations for various types of studies

In our lab, we run a large variety of studies, and each will have different "requirements" in a remote testing situation. For example, conducting a perception study based on reaction times (RT) will involve more safeguards than conducting a survey that does not involve any sound. Below is an attempt at providing a decision matrix for which recommendations and adjustments are needed for which type of study and which conditions are suitable for inclusion of the dataset in the analysis. This list is subject to change as we test and analyze more – feedback is welcome! (idarcy@indiana.edu)

Study / **Type of modification needed**	Survey (no sound)	Survey with sound (ratings…)	Perception – no RT	Perception – with RT	Visual (no sound) task	Production
Zoom meeting room	No - if it can be done alone	Maybe	Yes	Yes	Yes	Yes
Distraction check	No	Maybe	Yes	Critical	Yes	Maybe
Environment check	No	Yes/Maybe	Yes	Yes	Yes	Yes
OS check	No	No	Yes	Yes	Yes	No/Maybe
Requests for optimizing browser	No	Maybe	Yes	Critical	Critical	No/Maybe
Sound check	No	Yes	Critical	Critical	No	Yes if sound or low noise is essential
Headphone/ earbud required	No	Yes	Yes	Yes	No	No
Acceptable conditions for inclusion of dataset						
Using loudspeaker ok?	n. A.	Yes (sound check can be added for verification)	No	No	n. A.	Yes?
Using smart phone ok?	Yes	Yes? (with earbuds / headphones)	No	No	It depends on the task	Yes
Using tablet / touch screen ok?	Yes	Yes	Yes (if not messy)	No	It depends on the task	Yes
In Public/on the go ok?	Yes	No	No	No	No	No
Maximal recommended length	tbd	tbd	tbd	tbd	tbd	tbd

Specifics for Perception studies:

- Include in recruitment materials that the experiment requires headphones (for perception)
- Remind participants to put on their headphone before beginning the experiment
- Very specific to some perception studies:
  - It is also possible to do headphone checks to verify that people have headphones for instance. This can be very important in case of dichotic listening experiments, or in case of some designs where we need to send one sound file in one ear and another sound file in the other ear channel (like in auditory stroop experiments). For these experiments, participants have to have headphones on – they can't do it with loudspeakers. Headphone check : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5693749/

Specifics for Perception studies with RT:

- **Add an RT Baseline task:**
  Everybody will have a different keyboard, a different computer, a different setup of software. A lot of different things running in the background. Depending on the brand, there can be a lot of variability in the RT just due to the keyboard itself. A lot of added variability will also depend on the computer configuration. So that's a really important issue to try to get a handle on.
  We can alleviate some of this by not looking at absolute reaction times, but just rather **z-scoring** everyone to their own individual average RT, like essentially normalizing it within individual. We can also include a very brief task that they can do to just measure their absolute fastest reaction time. The task can be very simple and very short (< 2mins), and focus only on speed, such as "push spacebar when you see the red box appear on the screen". We are currently developing such a "baseline task" for the lab.
  We are currently considering either a picture distinction/choice, or an auditory baseline with left/right.
- **Very important: having a baseline and definitely don't use absolute RT values; comparison within participant across conditions is very important.** Obtaining the baseline could also help figuring out outliers because of equipment issues.
  - See https://www.psychopy.org/timing/2020/table3.html for interesting comparisons, and also the paper preprint available there. See also https://danluu.com/keyboard-latency/

Specifics for Production studies:

- We have brainstormed some ways to ensure a high-quality setup for production studies remotely – and we have experimented with various options.
  - **Smartphone tool:** Participants can use their smartphone if they own one, and the audiorecorder smartphone app. We've had no issues with asking them to download the app and record their speech as .wav format. Also if they have a small microphone, the sound quality is usually sufficient for our purposes.

- **Other tools are Praat and Audacity.** Both are free software that participants can download easily. We've had good experiences with them using either, but Audacity seems easier to handle for people.
    - The buffer limit (memory limit) in Praat can pose a serious problem for longer recording sessions! One possibility is to limit the sampling size to 16000 which is perfectly fine for speech – but this might be difficult to have participants do remotely.
    - The downside is that they need some sort of headset with microphone (the built in microphone of a laptop, for instance, would produce too noisy recordings). Another issue to consider is the echo. It's best if they can be in a room with lots of furniture and textiles (like a bedroom). A large empty room will probably generate echo (hollow sounding recording).
- **Procedure:** The experimenter provides the set of materials either as sounds to listen to (such as a delayed repetition task), or as a written set of materials. Powerpoint slides or a word document can be presented to the participant in a specific order, for example through sharing screen functions (through Zoom or Skype or similar apps). For recorded materials, a timed Powerpoint slideshow can be prepared and sent to the participant. They play the slideshow and record when prompts appear in the slideshow. Other recommendations about the use of sound (as above) apply in this case.
- **Instructions for file saving:** It is important to write a good set of instructions for the participant, also explaining where to put the file after it is recorded. It is also possible to give them a shareable link for Box – if this is available to you. Other cloud storage options are also possible.

Specifics for Survey studies:

- Optimize any survey for smartphone, especially if this is a survey-only study.
- If surveys are used together with a perception study, this is less crucial

How to collect handwritten responses from people:

- Vocabulary knowledge or familiarity with specific words is often an integral part of our research. Many tasks can be administered fine through a word document where people type their answers, but sometimes, it is necessary to obtain answers written by hand (for example when predictive typing would defeat the purpose of the task).
    - We have successfully done the following: we give participants a numbered list of words in English, asking them to translate in the L2 (in this case: Japanese), with the numbers, on a piece of paper. They then either took a photo of the piece of paper and sent it / uploaded it, or they held the paper in front of the screen and the experimenter took a screenshot of the piece of paper.
    - Another option, for example if the task requires participants to underline/circle/mark words in any way (e.g., for accent marking or other annotation like word stress…), is to provide a pdf document. The participants can use the "annotation tool" in pdf to mark

accents or circle things. Then, they can also send it/ upload it, or they can share their screen and you take a screenshot.

- We sometimes use ready-made vocabulary tests (like the X-lex/Y-lex, Meara 2005) which are installed on the testing computer. This is not likely to work in a remote testing situation, since they would need to be installed remotely. We have not yet programmed anything in Psychopy3. Right now, our best bet would be to look online for other vocabulary tasks that can be taken online and for which there is a record of performance (scores…). There are a few out there, it would be good to compile a list of resources here.

Conclusion

This is an exciting time for all of us, and learning to test online has the potential to transform our research. It can open new avenues for testing that we may not have ventured in before. But at the same time, it is essential that we do it well to ensure the highest quality research for the field. There will be difficulties at first, but we will get better at it as we do it. We should not expect perfection from the beginning.