Perception and Language Attitudes

KATIE DRAGER
UNIVERSITY OF HAWAIʻI AT MĀNOA
Why perception?
Some experimental paradigms
Some considerations
Some pitfalls to avoid
Designing in Excel
Playing with PsychoPy
Why perception?

- It’s fascinating! But also…

- Social meaning is co-constructed by the speaker and the listener

- Listeners have important roles in sound change (see Ohala)

- Previous sociophonetic perception work has only scratched the surface…
Alright, I’m convinced. I want to run an experiment. Now what?

1. Articulate your specific research question
   a) broad question as a research program
   b) narrow that question down into manageable chunks

2. Identify the appropriate experimental paradigm
Response vs. Response Times

Analyze response/accuracy if:

- you are new to running experiments
- you don’t have access to a button box and software that is accurate (e.g., EPrime, DirectRT, PsychoPy)
- you anticipate high error rates on your task
Matching research questions to a paradigm

- What social characteristics are attributed to Dialect X? What variables affect the perception of Social Category Y?
  → classification/categorization task or a rating task (matched guise)

- Can social information affect how a sound is perceived?
  → identification task or a lexical decision task

- Does a listener have positive or negative attitudes toward a certain linguistic variant or social group?
  → Implicit Association Task (IAT)
Classification:
what social characteristics are attributed to a speaker based on how they talk?

- occupation and SES (Labov 1966)

- ethnicity (Purnell et al. 1999)

- regional origin (Clopper & Pisoni 2004)

- vowels and perceived sexuality (Munson et al. 2006)
Modified matched-guise: what social characteristics are attributed to a phonetic variable?

- Resynthesize speech so that there are two versions of the same clip, each with a different phonetic variant

(Levon 2006, 2011; Campbell-Kibler 2007; Kirtley 2011)
‘... [s]lipped or [s]ome[th]ing...’

8 stimuli derived:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>[-FRONT]</td>
<td>[-SIB]</td>
<td>[-SHIFT]</td>
</tr>
<tr>
<td>b</td>
<td>[+FRONT]</td>
<td>[-SIB]</td>
<td>[-SHIFT]</td>
</tr>
<tr>
<td>c</td>
<td>[-FRONT]</td>
<td>[+SIB]</td>
<td>[-SHIFT]</td>
</tr>
<tr>
<td>d</td>
<td>[-FRONT]</td>
<td>[-SIB]</td>
<td>[+SHIFT]</td>
</tr>
<tr>
<td>e</td>
<td>[+FRONT]</td>
<td>[+SIB]</td>
<td>[-SHIFT]</td>
</tr>
<tr>
<td>f</td>
<td>[-FRONT]</td>
<td>[+SIB]</td>
<td>[+SHIFT]</td>
</tr>
<tr>
<td>g</td>
<td>[+FRONT]</td>
<td>[-SIB]</td>
<td>[+SHIFT]</td>
</tr>
<tr>
<td>h</td>
<td>[+FRONT]</td>
<td>[+SIB]</td>
<td>[+SHIFT]</td>
</tr>
</tbody>
</table>
## Rating Task

Levon (2006)

### Affective Scales

<table>
<thead>
<tr>
<th></th>
<th>1 = extremely</th>
<th>4 = neutral</th>
<th>7 = extremely</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. generous</td>
<td>1 2 3 4 5 6 7</td>
<td>greedy</td>
<td></td>
</tr>
<tr>
<td>2. lazy</td>
<td>1 2 3 4 5 6 7</td>
<td>hardworking</td>
<td></td>
</tr>
<tr>
<td>3. prudish</td>
<td>1 2 3 4 5 6 7</td>
<td>promiscuous</td>
<td></td>
</tr>
<tr>
<td>4. effeminate</td>
<td>1 2 3 4 5 6 7</td>
<td>masculine</td>
<td></td>
</tr>
<tr>
<td>5. aloof</td>
<td>1 2 3 4 5 6 7</td>
<td>friendly</td>
<td></td>
</tr>
<tr>
<td>6. straight</td>
<td>1 2 3 4 5 6 7</td>
<td>gay</td>
<td></td>
</tr>
<tr>
<td>7. neat</td>
<td>1 2 3 4 5 6 7</td>
<td>messy</td>
<td></td>
</tr>
<tr>
<td>8. savvy</td>
<td>1 2 3 4 5 6 7</td>
<td>naive</td>
<td></td>
</tr>
<tr>
<td>9. kind</td>
<td>1 2 3 4 5 6 7</td>
<td>mean</td>
<td></td>
</tr>
<tr>
<td>10. genuine</td>
<td>1 2 3 4 5 6 7</td>
<td>fake</td>
<td></td>
</tr>
</tbody>
</table>
Scales based on well-established ratings from psychology

Alternatively: before a rating task or a forced-choice task, you might want to run the experiment with open-ended questions or focus groups
  - messy but that reflects the fact that social information is complicated
  - can investigate social meaning that could be missed using predetermined categories

But how to analyze the data?
  - lists
  - judge (dis)agreement (Campbell-Kibler 2007) → used to design rating task
  - tag clouds (Drager et al. 2011)
  - topic modeling (LDA) (Schnoebelen & Drager 2014)
83 listeners asked about each voice

1. What word would you use to describe this person’s **style**?

2. What are **three words** you would use to describe this person’s **personality**?

3. What do you think this person **looks like** (hair, build, clothes, etc)?

4. What **gender** would you guess this speaker identifies with?

5. Please provide a guess regarding this speaker’s **sexual orientation**.

6. What **ethnicity/ethnicities** would you guess this speaker identifies with?

7. Where do you think this speaker is from (please be as specific as possible)?

8. What high school do you think they went to?

9. Do you think you know or have met this person? If so, what is the person’s name?
Only need 2 columns: ID, Text

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Gender</th>
<th>Year</th>
<th>Ethnicity</th>
<th>Orientation</th>
<th>Hair Type</th>
<th>Personality Type</th>
<th>Other Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Kyle Gordan</td>
<td>female</td>
<td>20s</td>
<td>heterosexual African American</td>
<td>female</td>
<td>heterosexual African American</td>
<td>the kind that</td>
<td>tall, strong p 25</td>
</tr>
<tr>
<td>2</td>
<td>Kyle Gordan</td>
<td>female</td>
<td>20s</td>
<td>heterosexual Caucasian</td>
<td>female</td>
<td>heterosexual Caucasian</td>
<td>bubbly &amp; energetic</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Kyle Gordan</td>
<td>male</td>
<td>20s</td>
<td>heterosexual Pacific Islander</td>
<td>male</td>
<td>heterosexual Pacific Islander</td>
<td>gets nervous</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Kyle Gordan</td>
<td>female</td>
<td>30s</td>
<td>heterosexual Caucasian</td>
<td>female</td>
<td>heterosexual Caucasian</td>
<td>likes to read</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Kyle Gordan</td>
<td>female</td>
<td>30s</td>
<td>heterosexual Pacific Islander</td>
<td>female</td>
<td>heterosexual Pacific Islander</td>
<td>kind of small</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Kyle Gordan</td>
<td>male</td>
<td>20s</td>
<td>heterosexual Caucasian</td>
<td>male</td>
<td>heterosexual Caucasian</td>
<td>very sensitive</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Kyle Gordan</td>
<td>female</td>
<td>20s</td>
<td>heterosexual Caucasian</td>
<td>female</td>
<td>heterosexual Caucasian</td>
<td>has a clear sense</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Kyle Gordan</td>
<td>male</td>
<td>20s</td>
<td>heterosexual Caucasian</td>
<td>male</td>
<td>heterosexual Caucasian</td>
<td>bit of a downer</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Kyle Gordan</td>
<td>female</td>
<td>20s</td>
<td>heterosexual Caucasian</td>
<td>female</td>
<td>heterosexual Caucasian</td>
<td>excitable &amp; kind</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Kyle Gordan</td>
<td>male</td>
<td>20s</td>
<td>heterosexual Caucasian</td>
<td>male</td>
<td>heterosexual Caucasian</td>
<td>gets his kicks</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Kyle Gordan</td>
<td>female</td>
<td>30s</td>
<td>heterosexual Asian</td>
<td>female</td>
<td>heterosexual Asian</td>
<td>introspective &amp;</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Kyle Gordan</td>
<td>male</td>
<td>20s</td>
<td>heterosexual African American</td>
<td>male</td>
<td>heterosexual African American</td>
<td>tries too hard</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Kyle Gordan</td>
<td>female</td>
<td>20s</td>
<td>heterosexual Pacific Islander</td>
<td>female</td>
<td>heterosexual Pacific Islander</td>
<td>is funny when</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Gene</td>
<td>female</td>
<td>20s</td>
<td>heterosexual African American</td>
<td>female</td>
<td>heterosexual African American</td>
<td>tall, strong p 25</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Gene</td>
<td>female</td>
<td>20s</td>
<td>heterosexual Caucasian</td>
<td>female</td>
<td>heterosexual Caucasian</td>
<td>excitable, loud</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Gene</td>
<td>male</td>
<td>20s</td>
<td>heterosexual Caucasian</td>
<td>male</td>
<td>heterosexual Caucasian</td>
<td>shorter, bald</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Gene</td>
<td>female</td>
<td>20s</td>
<td>heterosexual African American</td>
<td>female</td>
<td>heterosexual African American</td>
<td>nice, interested</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Gene</td>
<td>female</td>
<td>18-20</td>
<td>heterosexual Caucasian</td>
<td>female</td>
<td>heterosexual Caucasian</td>
<td>black hair, shy</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Gene</td>
<td>male</td>
<td>17-20</td>
<td>heterosexual Caucasian</td>
<td>male</td>
<td>heterosexual Caucasian</td>
<td>curious, interested</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Gene</td>
<td>female</td>
<td>17-20</td>
<td>heterosexual Caucasian</td>
<td>female</td>
<td>heterosexual Caucasian</td>
<td>strong person</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Gene</td>
<td>male</td>
<td>30s</td>
<td>heterosexual Caucasian</td>
<td>male</td>
<td>heterosexual Caucasian</td>
<td>passive, caring</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Gene</td>
<td>male</td>
<td>20s</td>
<td>heterosexual Latino</td>
<td>male</td>
<td>heterosexual Latino</td>
<td>fast-paced, neat</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Gene</td>
<td>male</td>
<td>20s</td>
<td>heterosexual Caucasian</td>
<td>male</td>
<td>heterosexual Caucasian</td>
<td>strong person</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Gene</td>
<td>female</td>
<td>20s</td>
<td>heterosexual African American</td>
<td>female</td>
<td>heterosexual African American</td>
<td>tall female</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Gene</td>
<td>male</td>
<td>20s</td>
<td>heterosexual Caucasian</td>
<td>male</td>
<td>heterosexual Caucasian</td>
<td>laid back</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Gene</td>
<td>female</td>
<td>20s</td>
<td>heterosexual Polynesian</td>
<td>female</td>
<td>heterosexual Polynesian</td>
<td>shy passive</td>
<td></td>
</tr>
</tbody>
</table>
Empty topics

Topic 1

Topic 2
Let your computer do the work

- Lots of folks use Mallet and there are implementations in R
- We used the Stanford Topic Modeling Toolkit
5 social types for women

Types of attributes

- Physical descriptors
- Majority sexuality
- “Local”
- Two perspectives on energy levels
- Btw, men aren’t nice

Hair: curly, brown, dark
Skin: light, laid-back
Eyes: female, heterosexual, polynesian, hawaiian, haole
Tall
Nice
Lazy
5 social types for women

Types of attributes

- Physical descriptors
- Majority sexuality
- Asian and Filipina
- Socially engaged

- Female
- Straight
- Heterosexual
- Asian
- Japanese
- Chinese
- Filipino
- Height
- Average
- Short
- Long
- Hair
- Medium
- Dark
- Build
- Thin
- Casual
- Friendly
- Kind
- Fun
Now turning to experiments that investigate whether social information can affect how sounds are perceived...
Identification vs. Discrimination

- **Identification:**
  What word do you hear: *rarely* or *really*?

- **Discrimination:**
  Do these sound like the same word or different words?
Identification Tasks

- Real speech (Labov 1994; Thomas & Hay 2005; Hay et al. 2006b; Drager 2010)
  - words that have the same phoneme-level representation (e.g., mergers & homophones)

- Synthesized speech (Strand 1996; Johnson et al. 1999; Kaiser & Plichta 2009; Drager 2011; Jannedy & Weirich 2011)
  - continua between two sounds
  - voice quality (e.g., nasalization)
Identification Task with Photo Manipulation

- Different photos are paired with the same token, for different groups of participants

Example: Hay et al. (2006b) NEAR-SQUARE merger
NEAR/SQUARE Merger
-distinct speaker-
NEAR/SQUARE Merger

merged speaker

[Graph showing F1 vs F2 with data points for 'near' and 'square']
Prediction:

- Because the NEAR-SQUARE merger is led by young females from lower socioeconomic groups, if a listener believes a speaker is from this group (and therefore likely to merge) they may be less accurate at identifying tokens of NEAR and SQUARE.
Subjects rate characteristics about the people in the photographs

The test whether these ratings predict accuracy and, if they do, whether the relationship is consistent with sociophonetic trends in production

Results provide evidence that social info attributed to the person in the photo affects perception of a sound undergoing merger.
Implicit Association Task (IAT)

- Categorization task
- Measures reaction times
- Slower reaction times taken to indicate less alignment between concepts

see e.g., Campbell-Kibler 2012
Hawaii
good
USA
bad

Honolulu
<table>
<thead>
<tr>
<th>USA</th>
<th>Hawaii</th>
</tr>
</thead>
<tbody>
<tr>
<td>good</td>
<td>bad</td>
</tr>
</tbody>
</table>

Honolulu
Why IAT?

- Because it looks at reaction times, can investigate subtle biases that the participant wouldn’t self-report and that might not show up in response-based experimental paradigms.
Using reaction times

Advantages

- Tied with processing
- High error rates **not** required → wider range of stuff you can test

Warnings

- Must be run in experiment mode
- Regular testing and calibration with both software and hardware is required
- Control for accuracy
- Utmost control of stimuli
I have a research question and I know what experimental paradigm I’m going to use. What now?

Regardless of what experimental paradigm you use, there are certain things you should do before you start...
Take time to:

- keep it simple
  - start small (you can console yourself by promising to do follow-up experiments)
  - replicate previous experiments with your own “added twist”

- when narrowing down your research question, be specific about what questions the results can answer
  - what are the implications for every possible outcome?
  - what does a null result look like?
  - why is it worth asking? (frame within a larger research question)

- make sure you have the right stimuli, design, and participant group...
Checking your stimuli

» Normalize the stimuli for volume, background noise, amount of silence before and after the sound/word/utterance, or anything else that isn’t the factor you are testing.

» Before running your experiment, run a pilot on a select number of “subjects”.
  » for experiments testing accuracy, make sure you are getting incorrect responses some of the time
  » can you detect any unanticipated response biases?
  » any hick-ups in the programming, the stimuli, or design?

» If using synthesized tokens, you may need to test for naturalness.

» When using a large number of different words or sentences (e.g., IAT), you may need to conduct a norming study.
Tips for a naturalness experiment

- Tell people you want their help to see whether they can tell which tokens are synthesized.

- Rate how natural the words sound.

- Practice/training task:
  - 100% natural tokens
  - some very obviously synthesized tokens

- Include
  - all test items
  - all fillers
  - extra tokens so you can “top up”
  - optional: natural words and obviously synthesized ones
Tips for a norming experiment

- Use all potential test items and fillers (plus some items that you may need to use to replace any rejected stims)

- Ask participants to rate the items according to the social factor you are testing.
  Is this word more male, female, or neutral?
  - bank
  - leisure
  - lipstick
  - wrench

- Use the responses to identify appropriate items for the main experiment
How many subjects should I get? How many test tokens do I need? How many fillers?

It depends on the design and how you expect responses to be...
How complicated is your design

- Many conditions/predicting factors
  =
  more subjects

- Within-subject vs. between-subject design
  (fewer)  (more)
Think ahead to your analysis

- More variation in responses = more subjects and test items required
  - open-ended questions need a large number of subjects if you plan to conduct statistical analysis
If you have a simple experiment with a binary, forced-choice task and there is a great deal of consistency across how different participants respond, you may only need 20 subjects.
Do I need fillers?

When fillers aren’t needed
- pilot data (sometimes)
- if you want to highlight the point of the task (e.g., can people identify beer vs. bare if their attention toward the merger is highlighted)

When fillers are needed
- if you want to disguise the purpose of the experiment
- if you need to counterbalance the experiment so that participants aren’t answering the same thing all the time or so they can have some easy answers
How many fillers do I need?

- Twice as many fillers as test items is usually a safe bet, but it depends on your design.
- Have a look at what previous studies have done
How many voices do I need?

It depends on the design and what you are testing...
Number of voices

- **Identification tasks:**
  - can start with one voice
  - use voices with alternative social characteristics in follow-up experiments

- **Rating tasks:**
  - Need a minimum of two voices per social characteristic

- **Mind the time**
  - 4 voices (2 M, 2 F) x 20 sound clips x 2 guises x 6 questions = 960 test items…
  - …and you’d still need to add fillers!
Counterbalancing

- counterbalance: construct your experiment so that potential response biases are distributed evenly across the conditions/factors of interest
  - yes/no & order of response options
  - token frequency, phonotactic frequency, etc.

- blocks: arranging of stimuli in groups (usually based on similarity of stimuli)
  - blocking by gender or voice: remember to counterbalance
Randomization *

- randomization: different stimulus order every time the experiment is run
- pseudo-randomization: the order approximates random but there are strict constraints on the order
- multiple conditions: the order of tokens (and only the order) varies across conditions

* listed in descending order of desirability from most to least desirable
Remember:

- You want to control for everything except the specific factor that you are testing.

- While, in practice, this is impossible, there are some things you should avoid whenever possible...
Some pitfalls to avoid

- **effect of the experimenter**
  - when possible, use a single experimenter who is blind to the purpose of the study

- **unanticipated primes in the room**
  - e.g., posters
  - if not possible to control, at least keep constant across any & all conditions
Other Thoughts & Advice

- keep the experiments short
  - 15 minutes is great
  - if long, give subjects plenty of breaks

- don’t be tempted to look at the data too soon, but don’t wait until you’ve run 80 participants either
  - I like to look after running 10-15 subjects in each condition

- try to balance listeners’ social characteristics across the different conditions

- collect production data, too, if possible

- background information sheets
  - where subjects are from, their gender, age, L1(s), and whether they’ve taken linguistics courses

- online vs. in-person

- practice rounds
Questions so far?

- After this, we’ll be designing an experiment
PsychoPy

- Open-source
- Platform independent
- Written in Python

- In our lab, consistently 40ms off

- Good for: forced-choice and rating tasks
- Not good for: open-ended answers

- Let’s have a look...
Thank you!

Big thank you to:

Laura Staum Casasanto,
Erez Levon,
and Rachel Schutz
for sharing their slides and sound files