Running web-based experiments

Judith Degen & Sebastian Schuster
LSA 2021 minicourse
Jan 7, 2021

Overview/schedule

1. 9:00-10:30: **Introduction**

   - How do crowd-sourced experiments work?
   - What kinds of experiments can be run online?
   - Example study
   - Using GitHub for research projects
   - Preregistration and open science

2. 10:30-10:40: **Break**

3. 10:40-12:30: **Tutorials**

   - Git and GitHub
   - Modifying an existing experiment
   - Preregistering an experiment
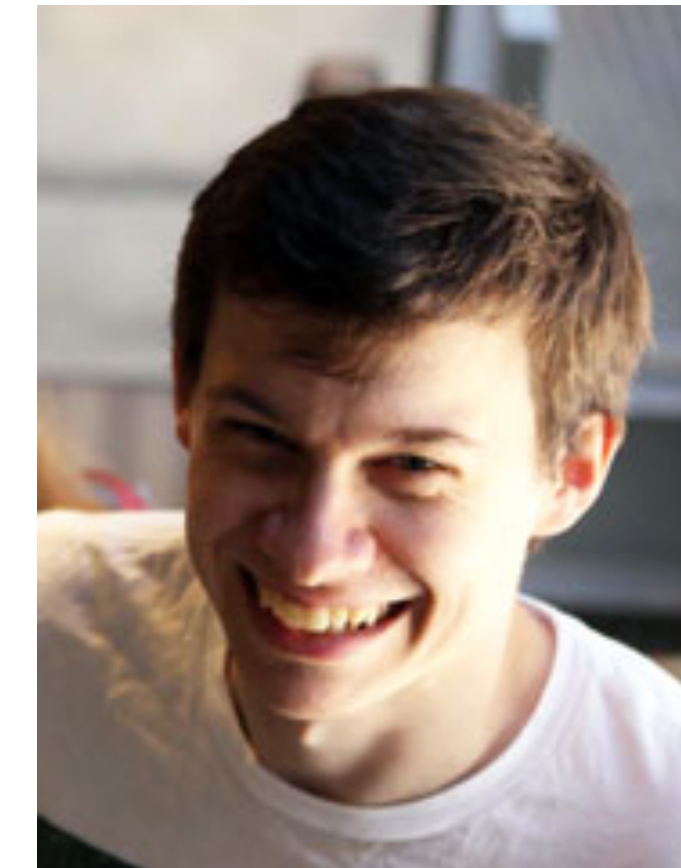
4. 12:30-1:00: **Lunch break**

5. 1:00-2:00: **Tutorials (continued)**

   - Posting the experiment
   - Testing the experiment
   - Downloading and visualizing the data

6. 2:00-2:15 **Final discussion and Q&A**

Judith Degen
jdegen@stanford.edu

Sebastian Schuster
schuster@nyu.edu

Mini course website:
https://sebschu.github.io/lsa-web-based-experiments/

Companion paper:
Degen, J. & Tonhauser, J. (to appear). Managing web experiments for psycholinguistics: An example from experimental semantics/ pragmatics. In *Open Handbook of Linguistic Data Management*.

Why run web-based experiments?

- more efficient data collection than in the lab

  - faster

  - cheaper

- larger & more diverse participant population (not just college undergrads)

  Buhrmester et al 2011

- …the pandemic

Concerns about web-based experimentation

- participants don't pay attention over the web   **attention checks**

- no control over participant population   **demographic parameters, "qualifications"**

- data is noisy   **many classic findings have replicated**
                                                              remedies
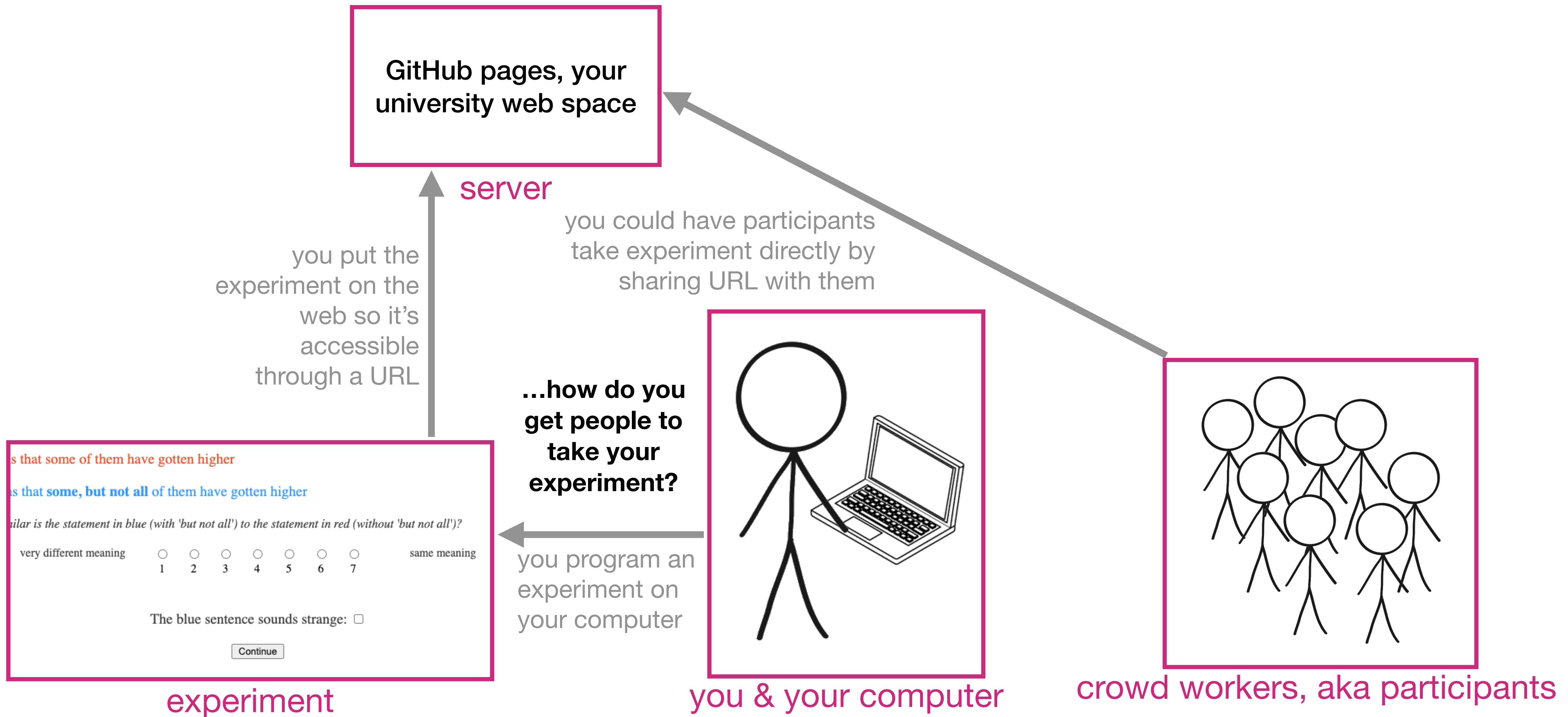
- difficult to learn

- too many tools/choices   **that's why we're here!**
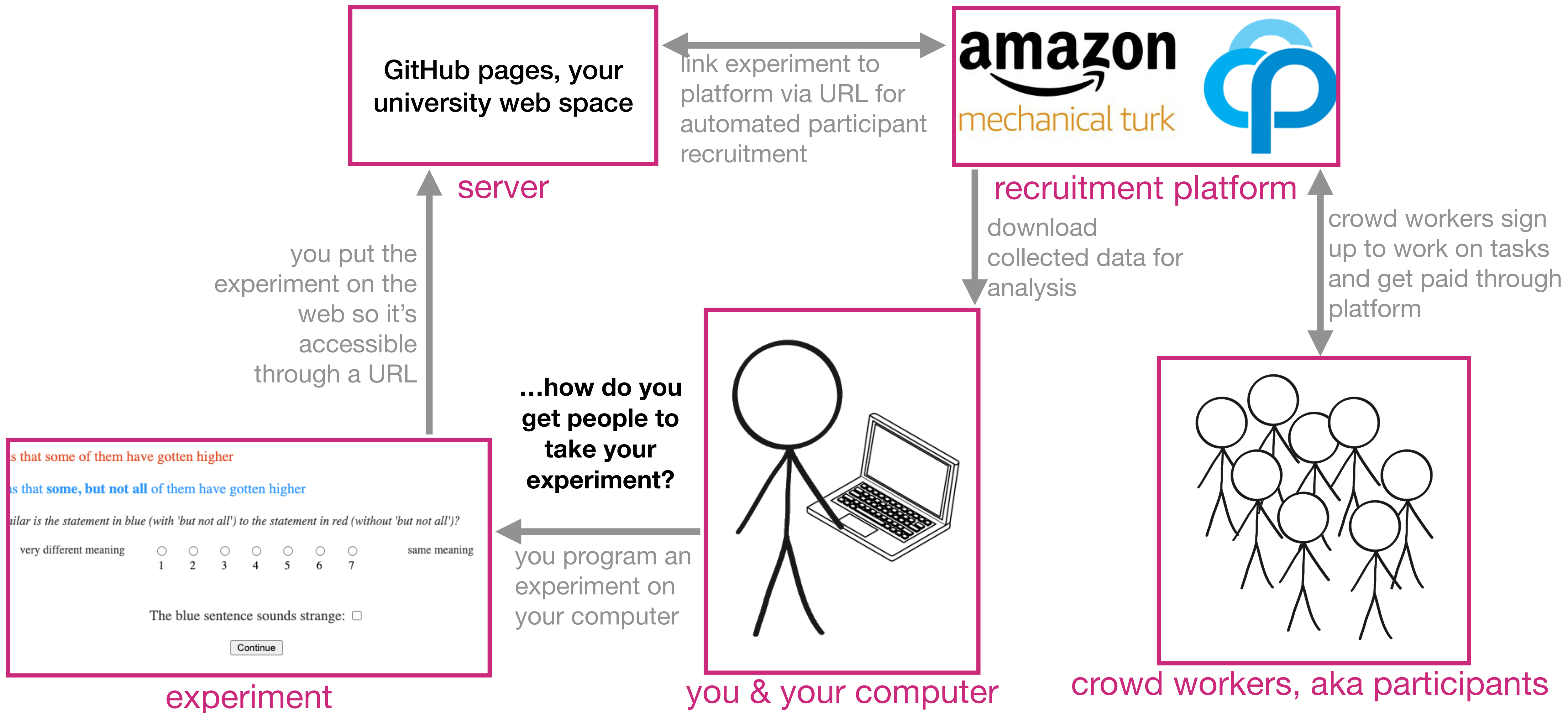
- don't know how to code

                                                        …additional concerns?

How do crowd-sourced experiments work?

How do web-based experiments work?

GitHub pages, your university web space

server

you could have participants take experiment directly by sharing URL with them

you put the experiment on the web so it's accessible through a URL

...how do you get people to take your experiment?

you program an experiment on your computer

you & your computer

crowd workers, aka participants

s that some of them have gotten higher

s that **some, but not all** of them have gotten higher

ilar is the statement in blue (with 'but not all') to the statement in red (without 'but not all')?

very different meaning   ○ ○ ○ ○ ○ ○ ○   same meaning
                          1   2   3   4   5   6   7

The blue sentence sounds strange: ☐

Continue

experiment

How do web-based experiments work?

GitHub pages, your university web space

server

link experiment to platform via URL for automated participant recruitment

amazon mechanical turk

recruitment platform

you put the experiment on the web so it's accessible through a URL

download collected data for analysis

crowd workers sign up to work on tasks and get paid through platform

...how do you get people to take your experiment?

s that some of them have gotten higher

s that **some, but not all** of them have gotten higher

ilar is the statement in blue (with 'but not all') to the statement in red (without 'but not all')?

very different meaning ○ ○ ○ ○ ○ ○ ○ same meaning
　　　　　　　　　　1　2　3　4　5　6　7

The blue sentence sounds strange: ☐

Continue

experiment

you program an experiment on your computer

you & your computer

crowd workers, aka participants

Workflow

Step	Tool	Time
1. create organized, version-controlled project repository	git GitHub	mins

Workflow

Step	Tool	Time
1. creat...	git GitHub	mins

Workflow

Step		Tool	Time
1. create organized, version-controlled project repository		git / GitHub	mins
2. program experiment and **test** in local browser		HTML / JS / CSS	hrs
3. put experiment on server and **test** again		scp/GitHub pages/…	secs + mins
4. preregister experiment		OSF	mins-hrs
5. link experiment to platform (and **test** again in Sandbox)		proliferate/ submiterator/…	secs + mins
6. run experiment on platform		amazon mechanical turk	hrs to days
7. download data from platform		proliferate/ submiterator/…	secs
8. pre-process, visualize, analyze data		R / RStudio	hrs
9. …publish!			…

Workflow

Step	Tool	Time
1. create organized, version-controlled project repository	git GitHub	mins
2. program experiment and **test** in local browser	HTML JS CSS	hrs
3. put experiment on server and **test** again	scp/GitHub pages/…	secs + mins
4. preregister experiment	OSF	mins-hrs
5. link experiment to platform (and **test** again in Sandbox)	proliferate/ submiterator/…	secs + mins
6. run experiment on platform	amazon mechanical turk	hrs to days
7. download data from platform	proliferate/ submiterator/…	secs
8. pre-process, visualize, analyze data	R	hrs
9. …publish!		…

What kinds of experiments can you run?

- categorical or continuous choices (eg truth-value judgment, acceptability)

- response time measures (eg self-paced reading, timed lexical decision)

- free-form input (eg for written production)

- spoken and signed production (possible but requires more setup)

- mouse-tracking

- even basic eye-tracking (example here: https://github.com/leylakursat/cohort_webgazer)

What kinds of experiments *can't/shouldn't* you run?

- brain imaging (EEG, MEG, fMRI)

- psychophysical studies requiring very temporally controlled stimulus presentation

- unclear: more complex eye-tracking setups thus far not validated

- unclear: special populatios (for children see https://lookit.mit.edu/)

Choices, choices

Platforms

- **Amazon's Mechanical Turk (US-based)**

  - large population of workers (~500,000, mostly US-based)

  - diverse, but nevertheless biased

    "workers are diverse but not representative of the populations they are drawn from, reflecting that Internet users differ systematically from non-Internet users. Workers tend to be younger (about 30 years old), overeducated, underemployed, less religious, and more liberal than the general population. Within the United States, Asians are overrepresented and Blacks and Hispanics are underrepresented relative to the population as a whole" Paolacci & Chandler 2014

  - many bots and bot-like workers

- **Prolific Academic (UK-based)** ← today

  - smaller than MTurk

  - higher data quality

  - more linguistic diversity

- **Others: TurkPrime, Daemo, Finding Five**

Experimental frameworks

- Stanford framework ← **today**

- psiTurk — maintained by Todd Gureckis

- jsPsych — maintained by Josh de Leeuw

- _magpie — maintained by Michael Franke

Tips and tricks, experimental setup

For participants' sake:

- in experiment:

  - always include consent/legal info

  - include thorough but concise instruc

InterActive Language Processing Lab at Stanford

In the following experiment, you'll see a series of visual scenes and be asked to complete sentences about them. The HIT should take around 5-6 minutes. Please pay attention, and answer carefully. Thank you!

Continue

LEGAL INFORMATION:
We invite you to participate in a research study on language production and comprehension. Your experimenter will ask you to do a linguistic task such as reading sentences or words, naming pictures or describing scenes, making up sentences of your own, or participating in a simple language game.

There are no risks or benefits of any kind involved in this study.

You will be paid for your participation at the posted rate.

If you have read this form and have decided to participate in this experiment, please understand your participation is voluntary and you have the right to withdraw your consent or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. You have the right to refuse to do particular tasks. Your individual privacy will be maintained in all published and written data resulting from the study. You may print this form for your records.

CONTACT INFORMATION:
If you have any questions, concerns or complaints about this research study, its procedures, risks and benefits, you should contact the Protocol Director Meghan Sumner (650) 723-4284.

If you are not satisfied with how this study is being conducted, or if you have any concerns, complaints, or general questions about the research or your rights as a participant, please contact the Stanford Institutional Review Board (IRB) to speak to someone independent of the research team at (650)-723-2480 or toll free at 1-866-680-2906. You can also write to the Stanford IRB, Stanford University, 3000 El Camino Real, Five Palo Alto Square, 4th Floor, Palo Alto, CA 94306 USA.

If you agree to participate, please proceed to the study tasks.

Tips and tricks, experimental setup

For participants' sake:

- in experiment:

  - always include consent/legal info

  - include thorough but concise instructions

  - include a progress bar (lowers dropout rates)

- on platform:

  - give your study a straightforward name (eg, "Language study")

  - include keywords that make it easy to find your study (eg, "language experiment study communication stanford fun cognitive science")

Tips and tricks, experimental setup

For your sake:

- in experiment:

    - to improve data quality:

        - include attention checks (eg, control trials with expected outcome)

        - include functionality to prevent experiment being taken repeatedly (UniqueTurker)

        - include functionality to identify bots/prevent them from taking experiment (eg, reCaptcha)

    - to not kick yourself later

        - always err on the side of recording too much rather than too little information

- on platform, to improve data quality

    - set participation parameters (eg, native language, percentage of previous work approved (>98%), participant location)

Tips and tricks, execution

- to improve data quality:

  - pay fairly (poorly paid studies -> unreliable data)   Buhrmester et al 2018

  - run studies on weekdays, in the morning

- to prevent worker frustration:

  - monitor email while experiment is running

  - Turkers communicate — sign up for Turkernation and/or Turkopticon account

Example study: variability and context-dependence of scalar inferences

Scalar implicatures in the wild

1. I like **some country music**.

   Inference? I like some, but not all, country music

2. It would certainly help them to appreciate **some of the things we have here**.

   Inference? …to appreciate some, but not all…

3. You sound like you have **some small ones** in the background.

   Inference? … some, but not all small ones…

Combining corpora & the web

1. extracted all 1390 utterances containing *some* from the Switchboard corpus of spoken American English

2. collected inference strength ratings for each item on Mechanical Turk (10 judgments per item)

Speaker A: i mean, they just have beautiful, beautiful homes and they have everything. the kids only wear
    name brand things to school and it's one of these things,

Speaker B: oh me. well that makes it hard for you, doesn't it.

Speaker A: well it does, you know. it really does because i'm a single mom and i have a thirteen year old
    now and uh, you know, it does.

Speaker B: oh, me.

Speaker A: i mean, we do it to a point but uh, not to where she feels different ,

Speaker B: yeah.

Speaker A:
        but some of them are very rich

---

but **some, but not all** of them are very rich

How similar is the statement with 'some, but not all' (green) to the statement with 'some' (red)?

Very different meaning ○ ○ ○ ○ ○ ○ ○ Same meaning
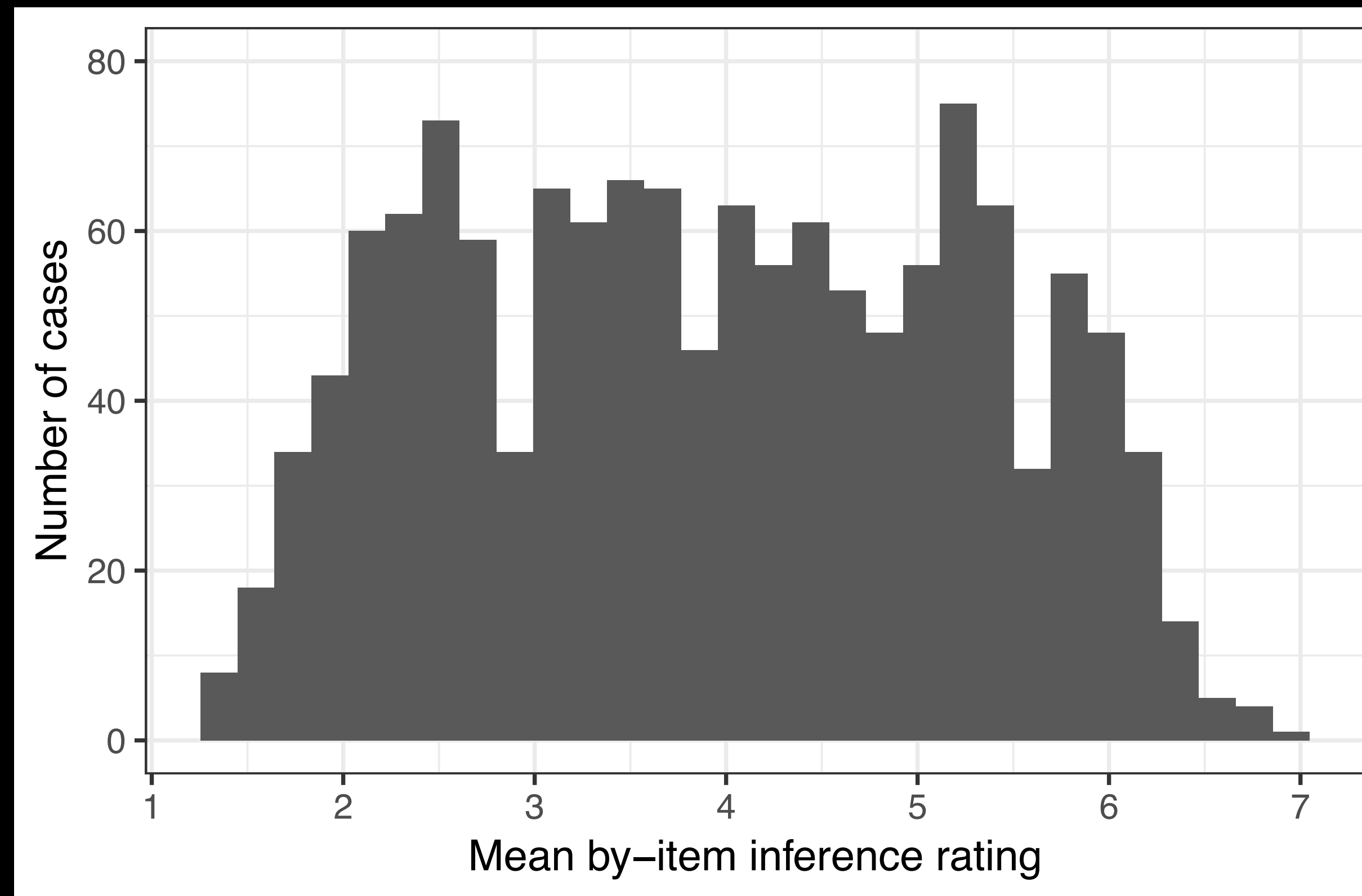
1   2   3   4   5   6   7

Continue

Variability in inference strength

large amount of variability in inference strength

Just noise?

Qualitative investigation

1. I like **some country music**.

6.9 **Inference?** I like some, but not all, country music

2. It would certainly help them to appreciate **some of the things we have here**.

4 **Inference?** …to appreciate some, but not all…

3. You sound like you have **some small ones** in the background.

1.5 **Inference?** … some, but not all small ones…

Stronger inferences…
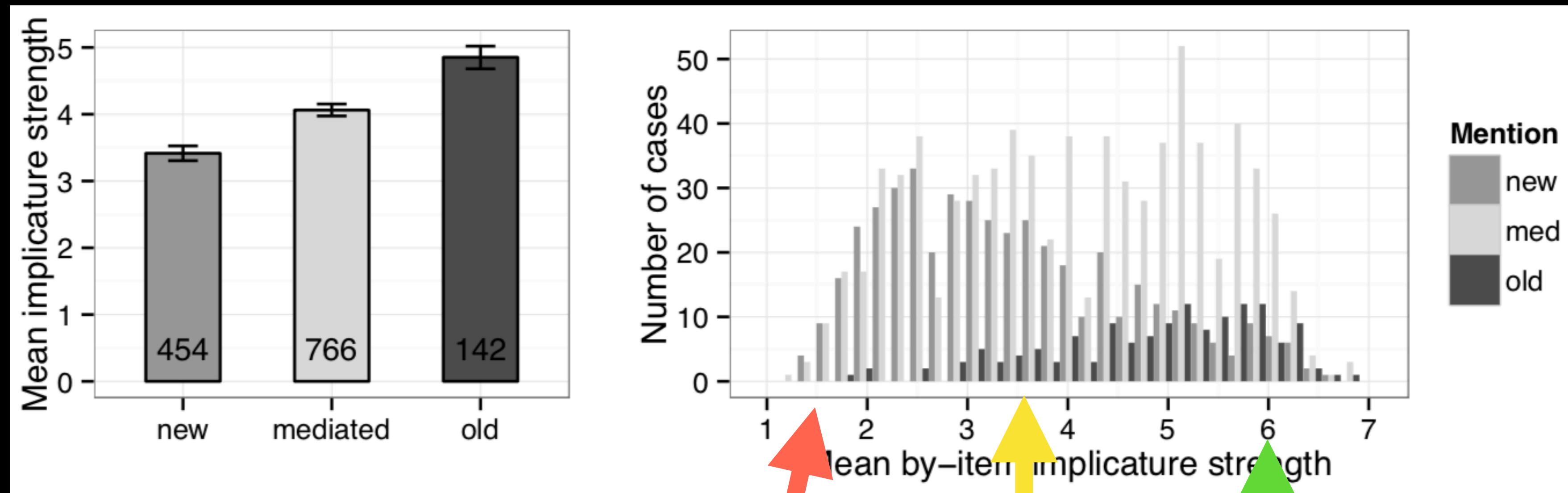
…with **partitive** *some*-NPs.
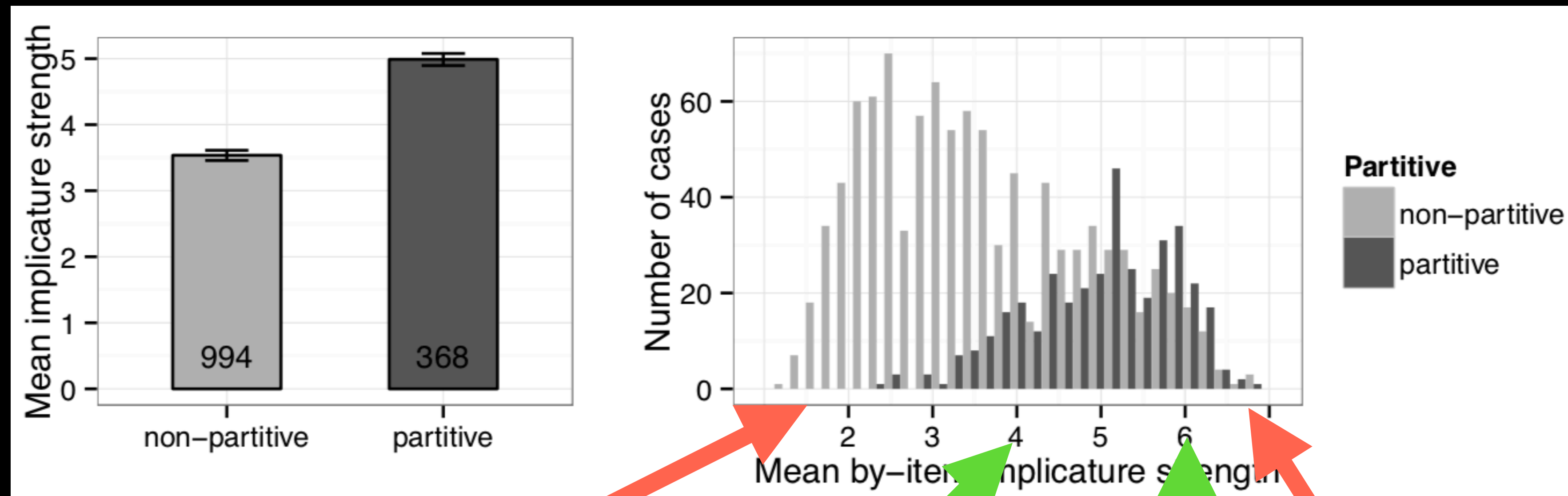
*I've seen **some of them** on repeat*

*It would certainly help them to appreciate **some of the things we have here**.*

*You sound like you have **some small ones** in the background.*

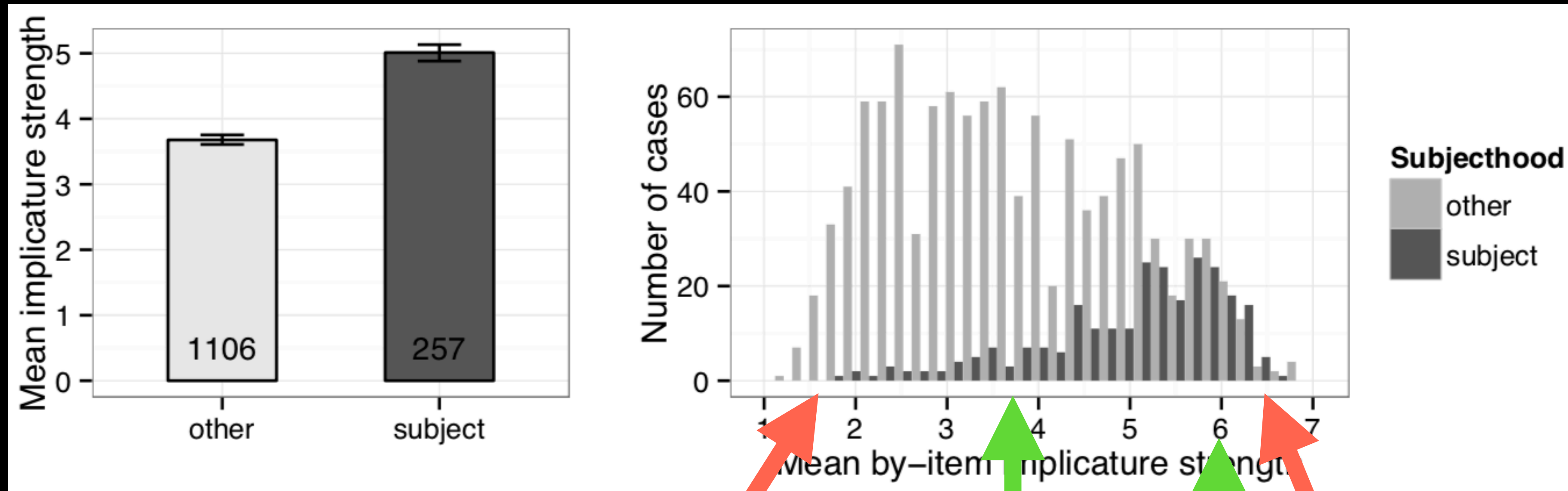*I like **some country music.***

Stronger inferences…

*I've seen **some of them** on repeats*

*We've got **some beets**.*

*That would take **some planning**.*

Stronger inferences…

…with *some*-NPs in **subject** position.

Some kids are really having it.

*Occasionally, **some ice skating** will come on.*

*That would take **some planning**.*

*I like **some country music.***

Just noise?

No. Variability in ratings is systematically predicted by syntactic, semantic, and pragmatic features of the linguistic context.

Next: open science

References

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?. *Perspectives on Psychological Science*, 6(1), 3-5.

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use. *Perspectives on Psychological Science*, 13(2), 149–154.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, 8(3), e57410.

Degen, J. & Tonhauser, J. (to appear). Managing web experiments for psycholinguistics: An example from experimental semantics/pragmatics. In *Open Handbook of Linguistic Data Management*.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23(3), 184-188.

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