

# Online Experiments for Language Scientists

Lecture 3: Grammaticality judgments

Kenny Smith

[kenny.smith@ed.ac.uk](mailto:kenny.smith@ed.ac.uk)

Finishing off the fire alarm lecture

# Con: Lack of control

In a normal lab study

- You interact with your participants when they arrive, and can see that they are indeed e.g. a human who speaks English natively
- They take part in a quiet, controlled lab environment on a modern machine that behaves in a known way
- You can monitor them as they participate, and they know this

With crowdsourced participants participating remotely, none of these things are true

- Consequently, experiments need to be designed to handle this

# Some ways to compensate for lack of control

- Add checks on who the participants are: native language checks, instruction comprehension checks, ...
- Add attention checks during the task, identify (and eject?) people who are not attending or who are responding randomly
- Can you make it easier to pay attention than not?
- Make the experiment short and fun! Most tasks on these platforms are pretty dull.



# Final note: Comparability with lab data

People often want to know if crowdsourced data is like lab data (i.e. do effects shown in the lab replicate online?)

- Lab data as a “gold standard” due to higher levels of control
- Or just because the effect you are interested in has only been shown in the lab

We’ll see numerous papers making direct comparisons, or replicating lab results with crowdsourced populations (e.g. the week 3 reading!)

OK, back on track

# What you will have read for today

(from [https://kennysmithed.github.io/oels2024/oels\\_reading\\_wk3.html](https://kennysmithed.github.io/oels2024/oels_reading_wk3.html) )

## Reading tasks for this week

Read:

- » Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, 43, 155-167.

As you read this paper and make note of any questions, criticisms or ideas it gives you, and I'll leave time in the Monday lecture slot so we can discuss these in class.

# Sprouse (2011)

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, 43, 155-167.

Compares undergrad lab and MTurk populations on grammatical acceptability judgment task

- Does the MTurk sample give similar judgments to lab population, despite reduced experimental control?



**Jon Sprouse**  
*NYU Abu Dhabi*



# Sample size, study duration etc

## Lab

- N=176
- Self-reported native speakers of English
- 96 sentences + practice items
- 30 minutes
- \$5 or course credit
- 3 months to collect

## MTurk

- N=176
- Self-reported native speakers of English
- 96 sentences + practice items
- No info on duration
- **\$3**
- 4 hours to collect

# Test items

Island effects (clear difference in ratings expected)

Grammatical (control): *What do you think that John bought?*

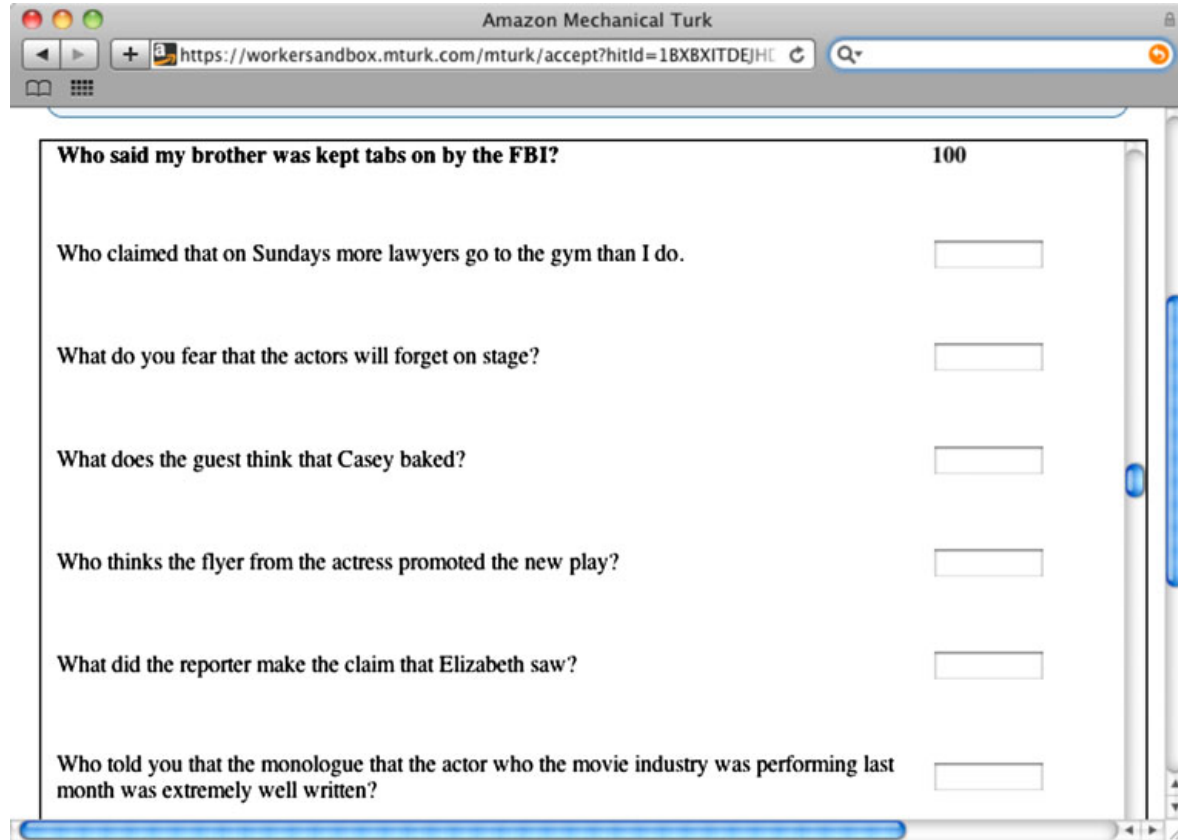
Ungrammatical (violation): *\* What do you wonder whether John bought?*

Illusions (smaller difference in ratings expected)

Clear ungrammatical (violation): *\* The slogan on the poster unsurprisingly were designed to get attention*

Ungrammatical? (illusion): *? The slogan on the posters unsurprisingly were designed to get attention*

# Task: magnitude estimation

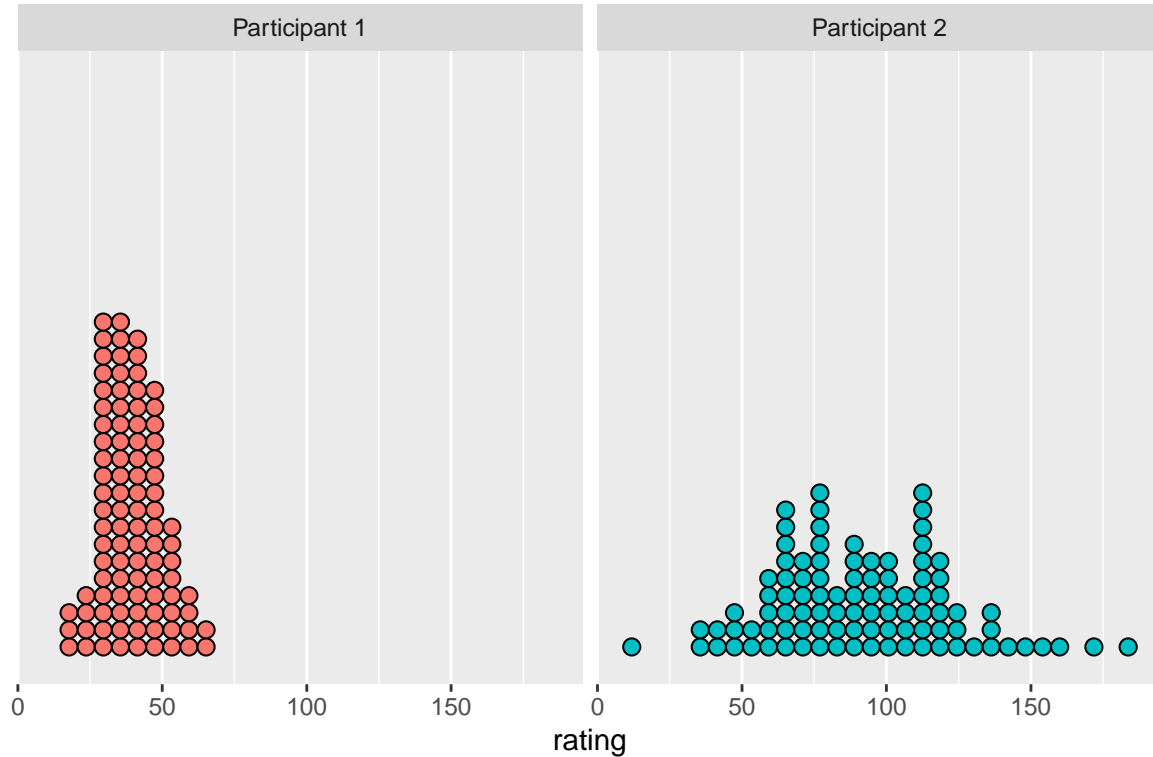


The screenshot shows a web browser window titled "Amazon Mechanical Turk" with the URL <https://workersandbox.mturk.com/mturk/accept?hitId=1BXBXITDEJHC>. The page contains a list of seven questions, each followed by a text input field. The first question, "Who said my brother was kept tabs on by the FBI?", has the number "100" next to it, indicating its assigned magnitude. The other questions are:

- Who claimed that on Sundays more lawyers go to the gym than I do.
- What do you fear that the actors will forget on stage?
- What does the guest think that Casey baked?
- Who thinks the flyer from the actress promoted the new play?
- What did the reporter make the claim that Elizabeth saw?
- Who told you that the monologue that the actor who the movie industry was performing last month was extremely well written?

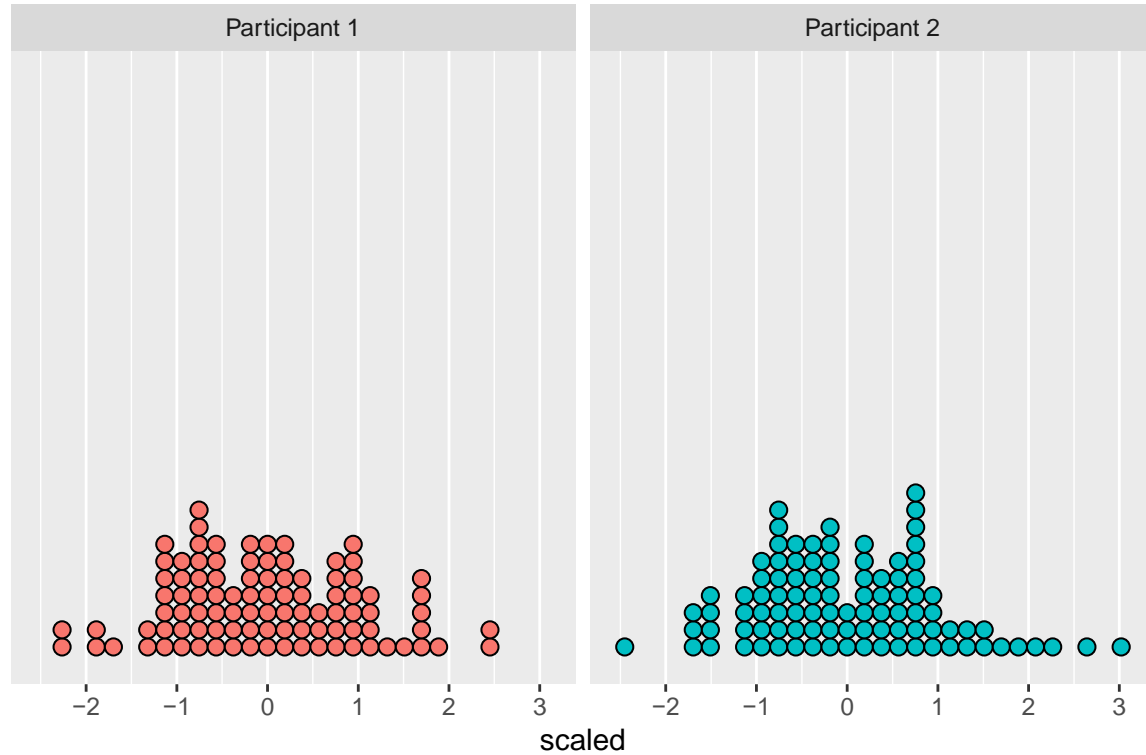
# Converting raw acceptability scores to z-scores

$$Z \text{ score} = \frac{\text{raw score} - \text{mean}}{\text{standard deviation}}$$



# Converting raw acceptability scores to z-scores

$$Z \text{ score} = \frac{\text{raw score} - \text{mean}}{\text{standard deviation}}$$



# Identifying outlier (inattentive?) participants

## (9) Examples of the Eight Conditions Chosen for the Rank Order Analysis

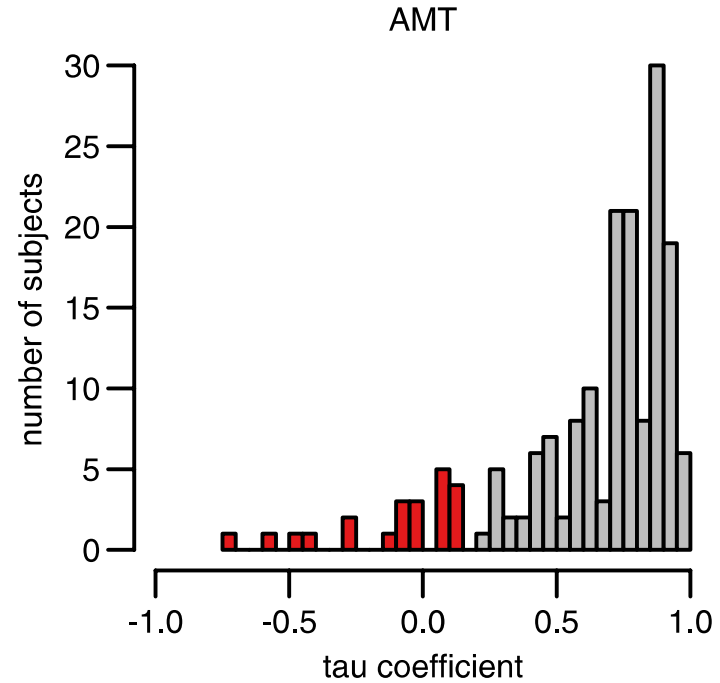
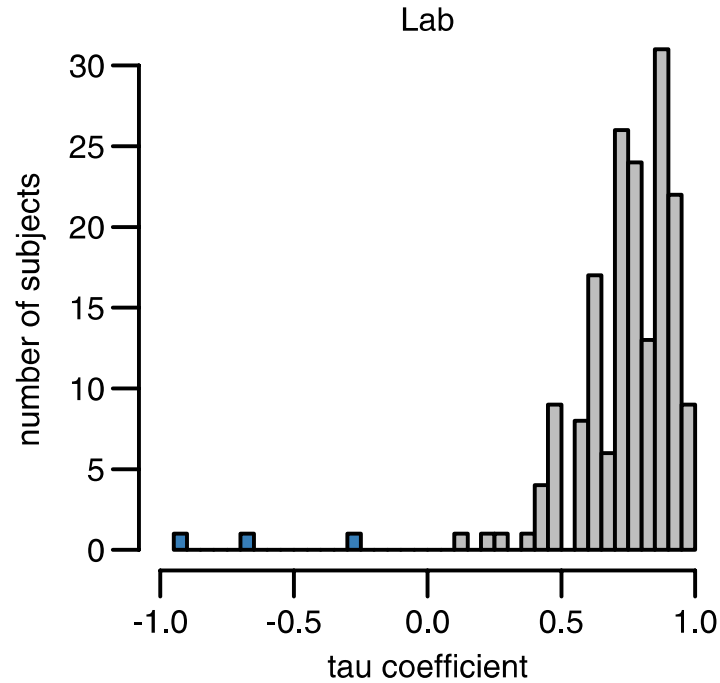
- a. What do you worry if the lawyer forgets at the office?
- b. What does the detective wonder whether Paul took?
- c. The slogan on the poster unsurprisingly were designed to get attention.
- d. The slogan on the posters unsurprisingly were designed to get attention.
- e. Who worries if the lawyer forgets his briefcase at the office?
- f. What does the detective think Paul took?
- g. Who made the claim that Amy stole the pizza?
- h. Who thinks Paul took the necklace?

Reliably low acceptability

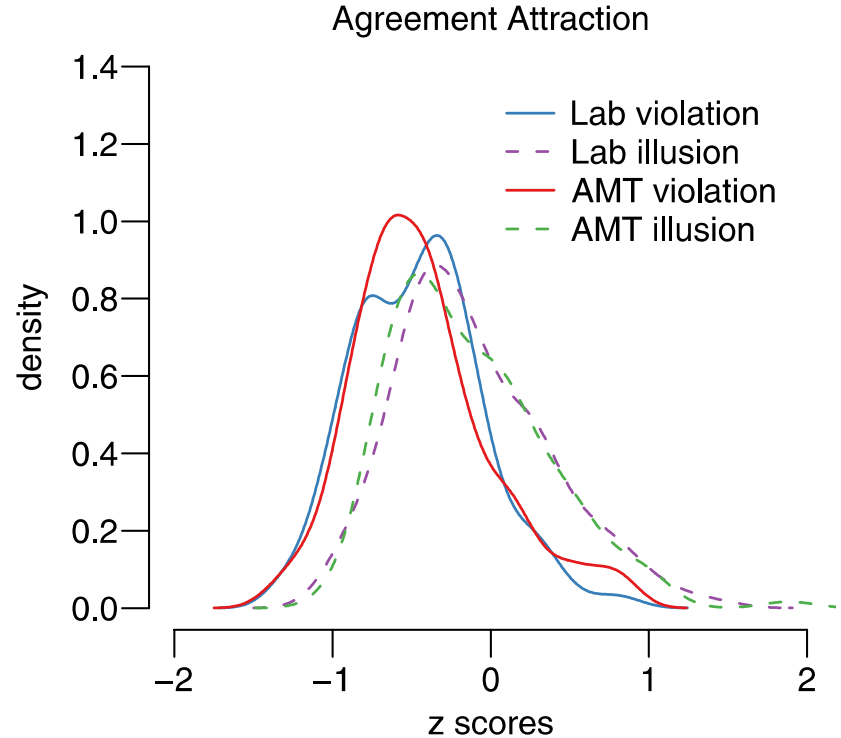
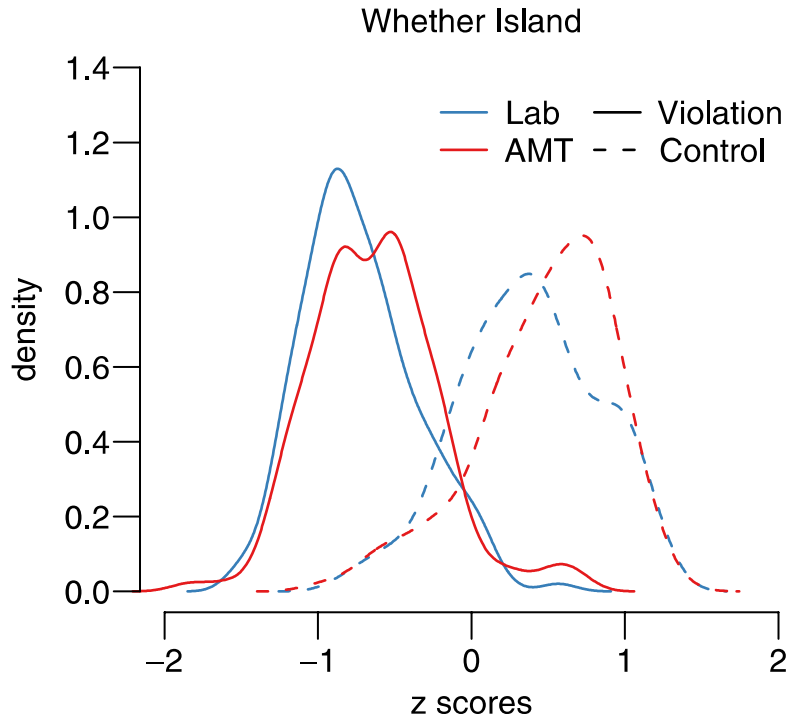


Reliably high acceptability

# Identifying outlier (inattentive?) participants

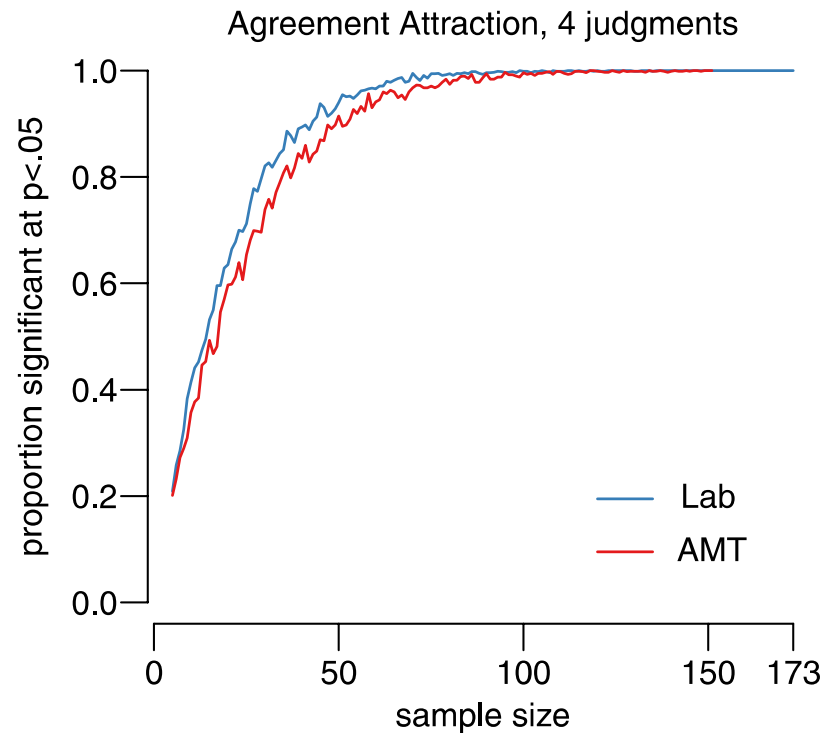
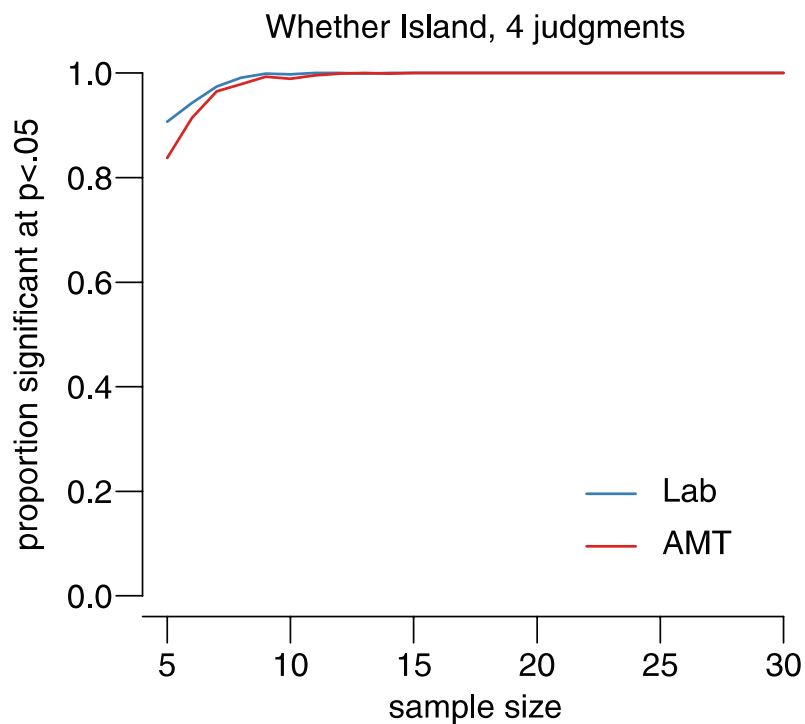


# Distribution of ratings

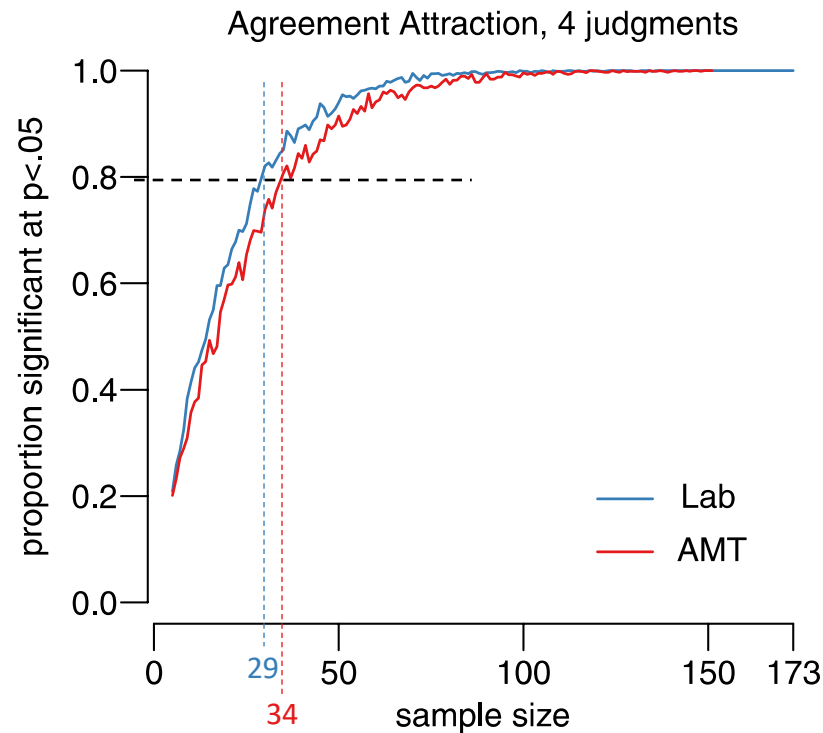
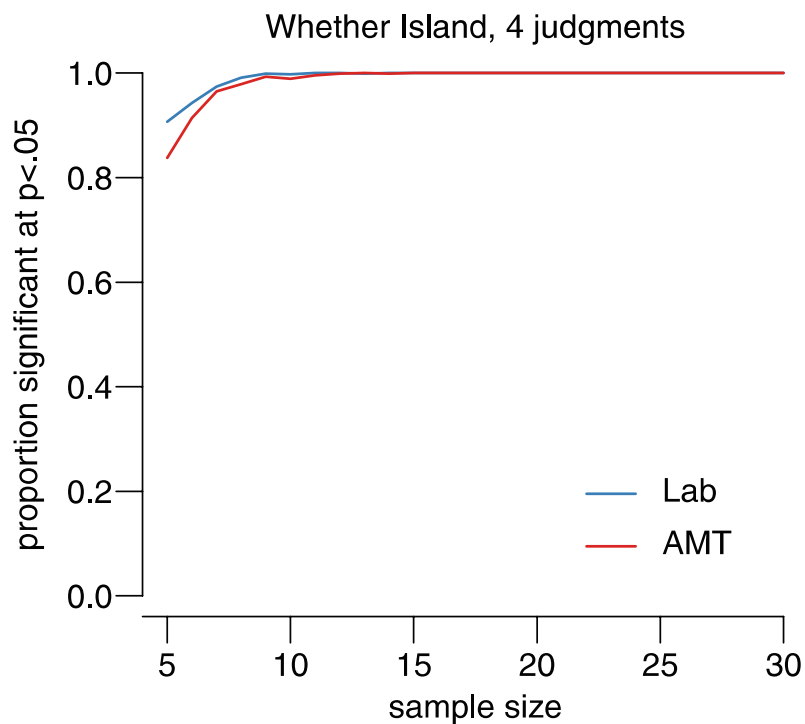




# Power calculations: how big does my sample size be to see the difference?



# Power calculations: how big does my sample size be to see the difference?



# Sprouse's conclusions

MTurk is suitable for collecting acceptability judgments

- Similar pattern of judgments in most places
- Small reduction in power (recommends increasing sample by 15%)
- Very fast

He also says some outdated stuff about limitations of online experiments re. presenting audio, collecting RTs etc – see my reading notes!

Time for Q&A/discussion on this week's reading

# Next up

## Wednesday: lab

- Our first proper experiment: grammaticality judgments
- I recommend taking a look at the materials in advance
- We give you the code, you mess with it

## Week 4

- Self-paced reading, do the reading before Monday's lecture!
- Week 4 lab will be a you-build-it-first lab!