

Online Experiments for Language Scientists

Lecture 3: Grammaticality judgments

Kenny Smith

kenny.smith@ed.ac.uk

What you will have read for today

(from https://kennysmithed.github.io/oels2023/oels_reading_wk4.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?*

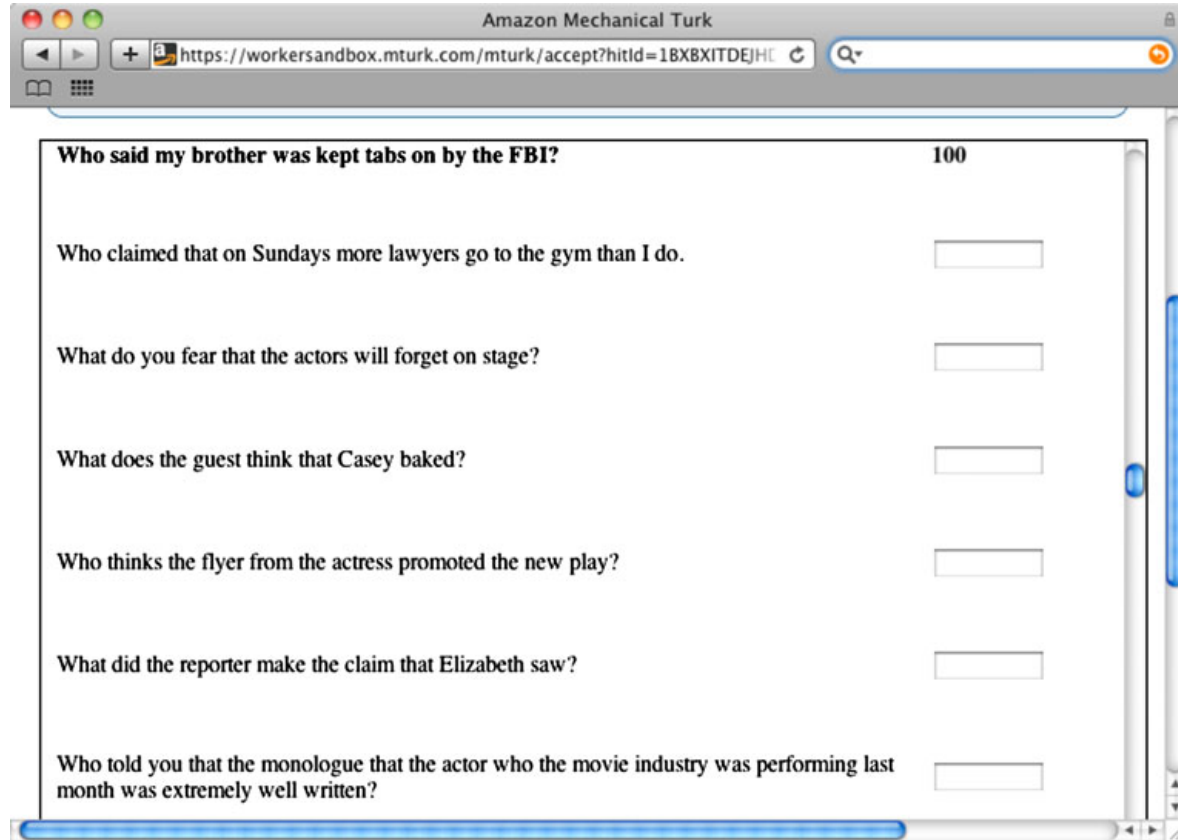
Ungrammatical (violation): ** What do you wonder whether John bought?*

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

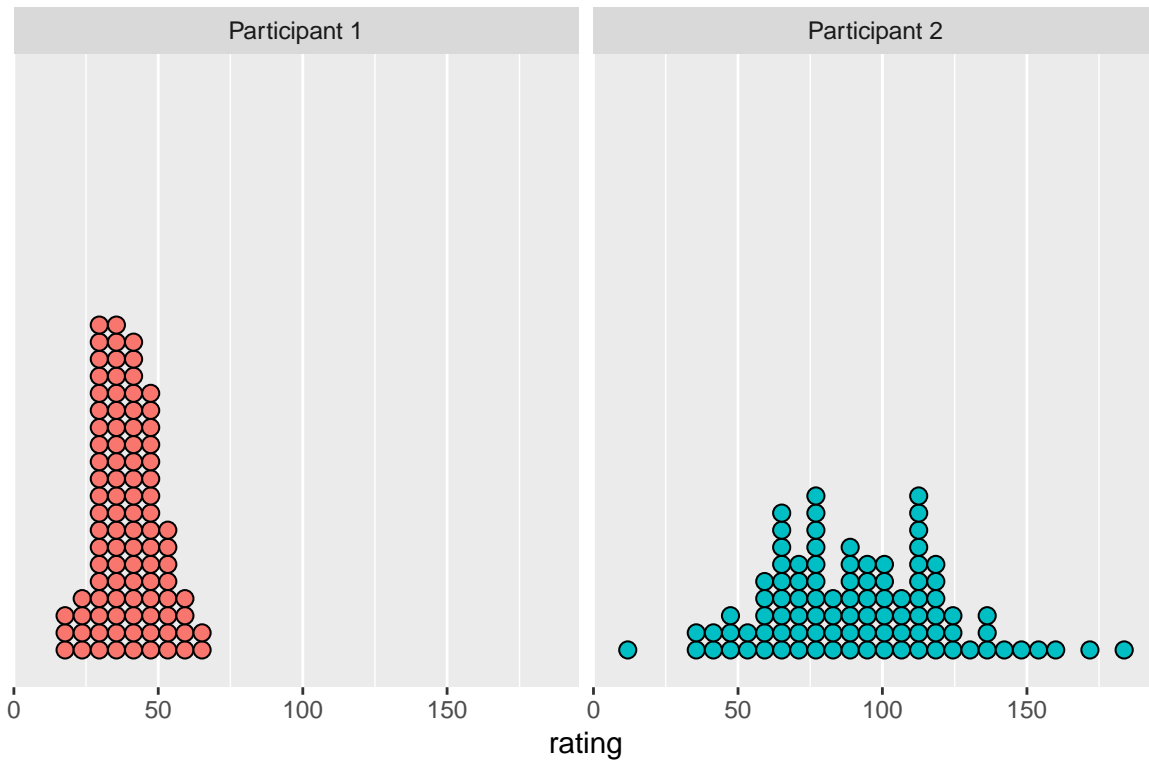


The screenshot shows a web browser window titled "Amazon Mechanical Turk". The address bar contains the URL "https://workersandbox.mturk.com/mturk/accept?hitId=1BXBXITDEJHC". The main content area displays a list of seven questions, each followed by a text input field. The first question, "Who said my brother was kept tabs on by the FBI?", is followed by the number "100". The other questions are followed by empty input boxes. A vertical scrollbar is visible on the right side of the content area.

Question	Response
Who said my brother was kept tabs on by the FBI?	100
Who claimed that on Sundays more lawyers go to the gym than I do.	<input type="text"/>
What do you fear that the actors will forget on stage?	<input type="text"/>
What does the guest think that Casey baked?	<input type="text"/>
Who thinks the flyer from the actress promoted the new play?	<input type="text"/>
What did the reporter make the claim that Elizabeth saw?	<input type="text"/>
Who told you that the monologue that the actor who the movie industry was performing last month was extremely well written?	<input type="text"/>

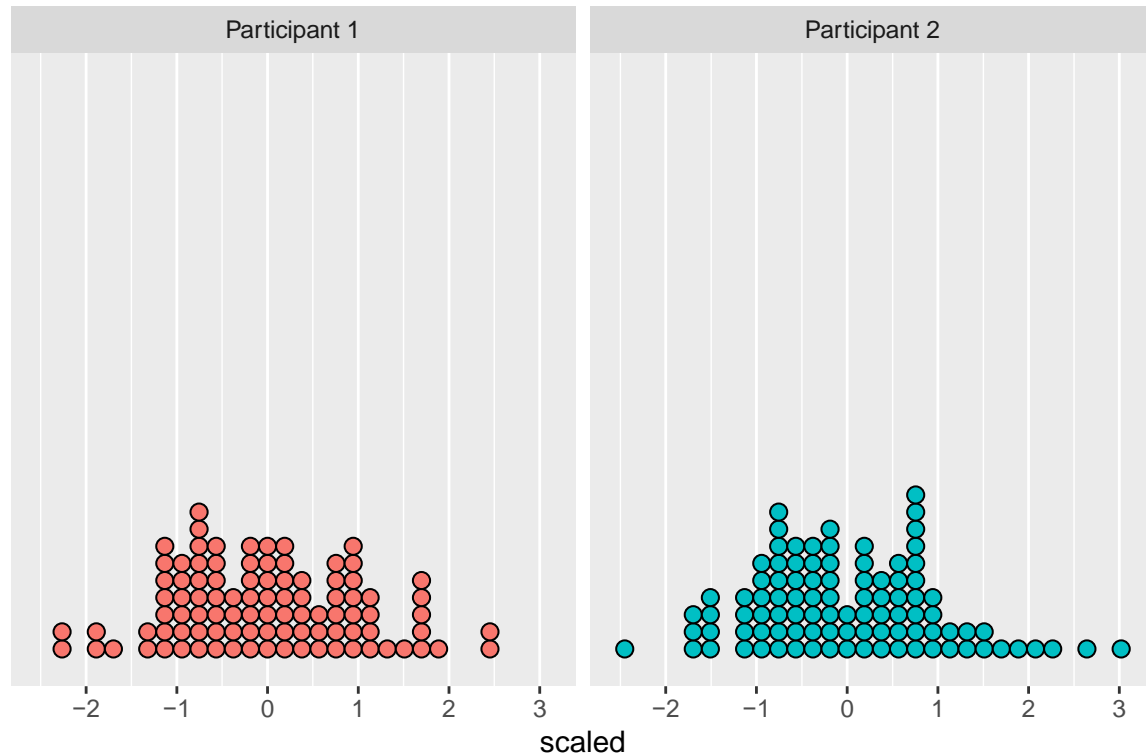
Converting raw acceptability scores to z-scores

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



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

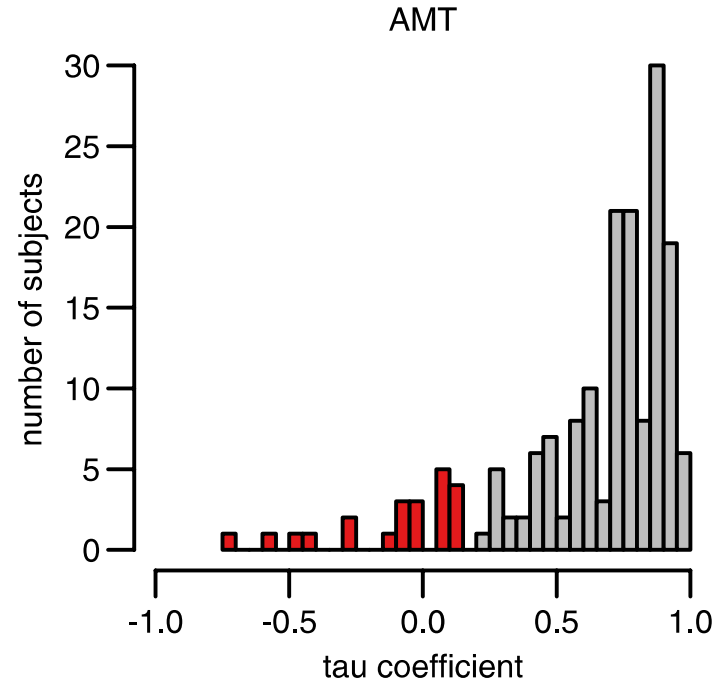
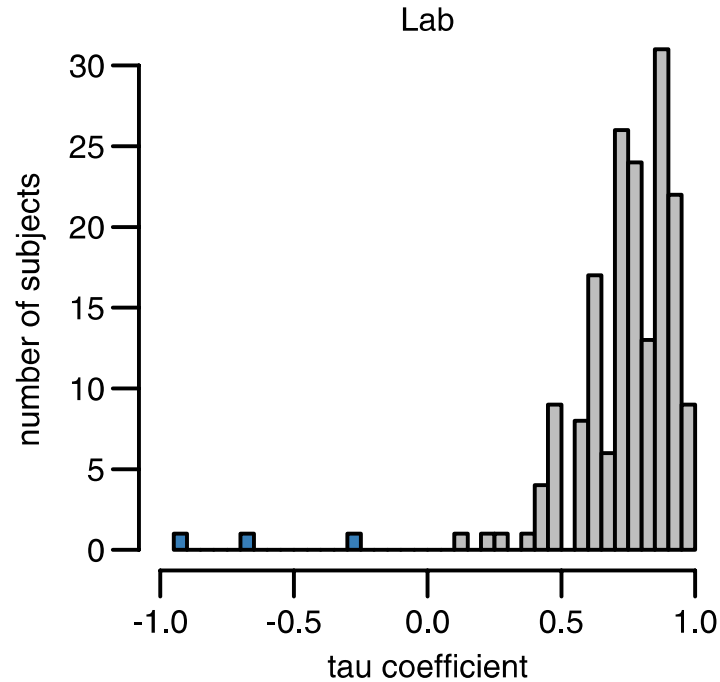
- a. What do you worry if the lawyer forgets at the office?
- b. What does the detective wonder whether Paul took?
- c. The slogan on the poster unsurprisingly were designed to get attention.
- d. The slogan on the posters unsurprisingly were designed to get attention.
- e. Who worries if the lawyer 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

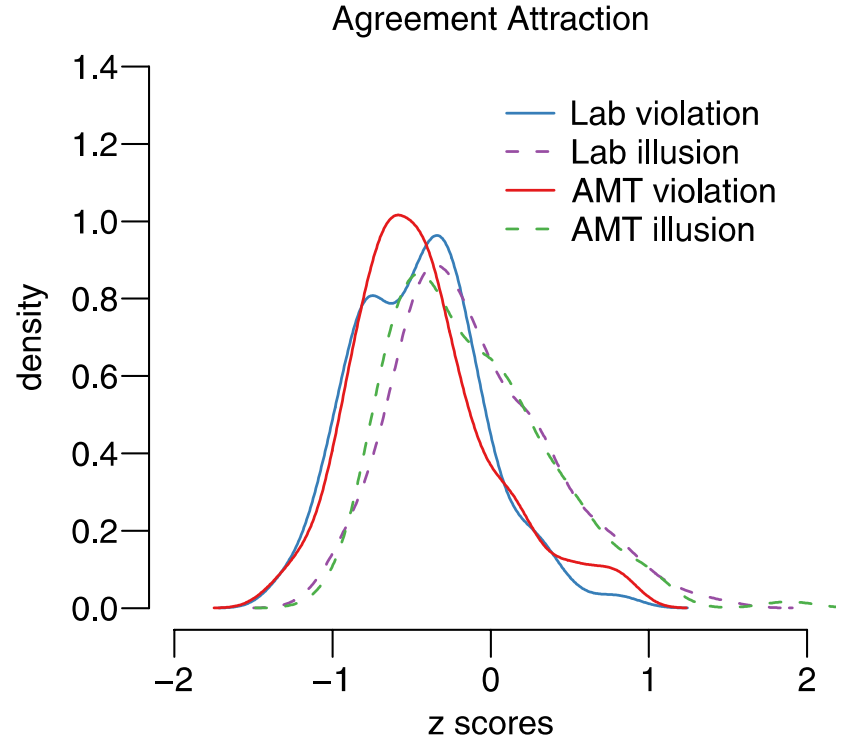
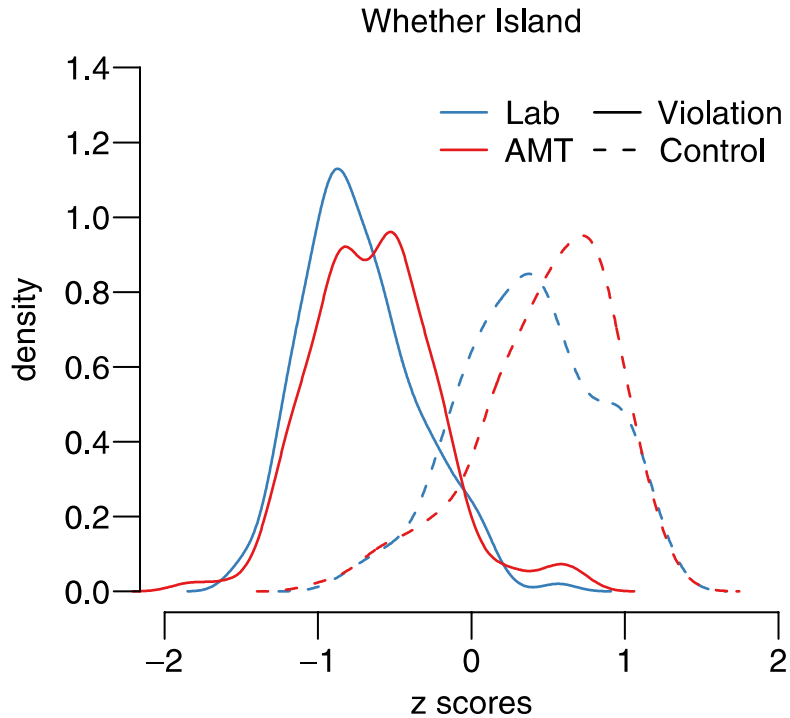


Reliably high acceptability

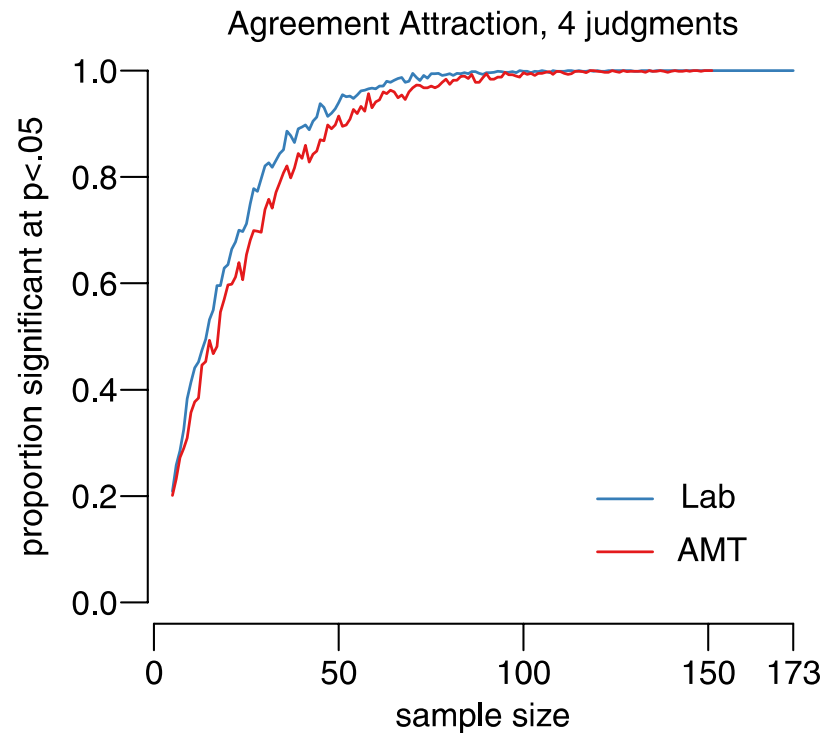
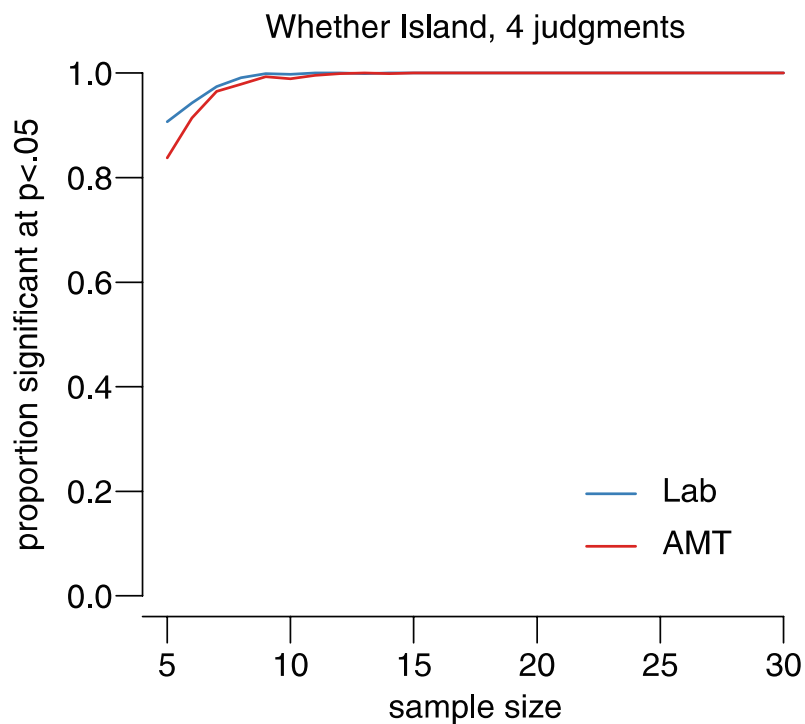
Identifying outlier (inattentive?) participants



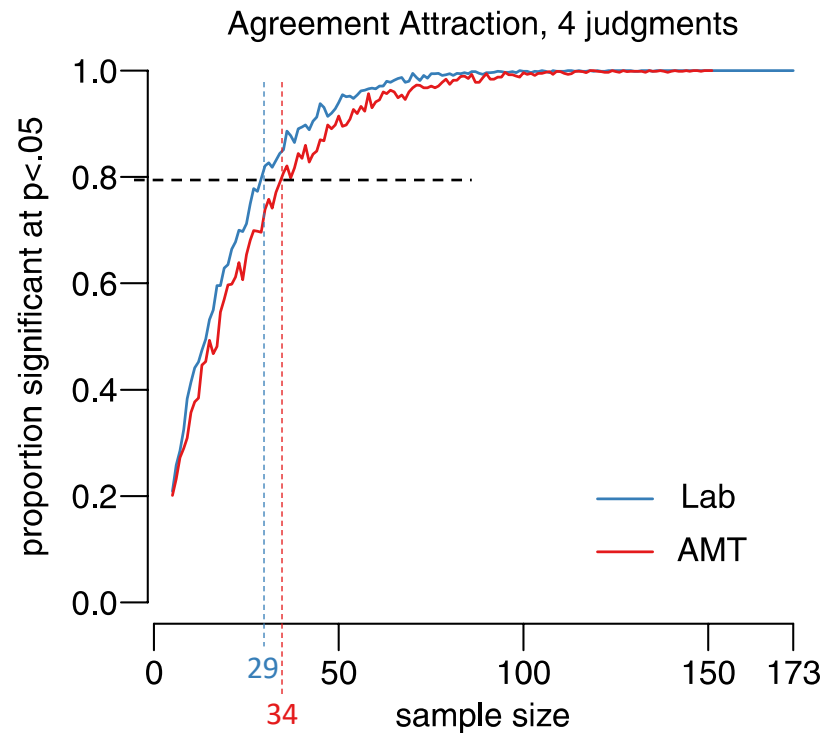
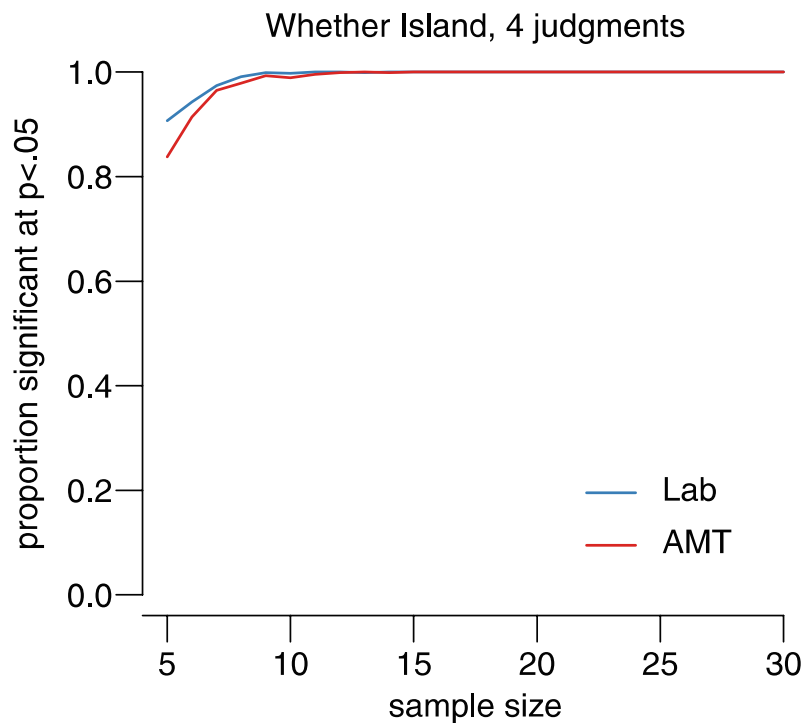
Distribution of ratings



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

Thursday: lab

- Our first proper experiment: grammaticality judgments

Week 5

- No lecture
- Catch-up lab

Week 6

- Lecture on self-paced reading, do the reading beforehand!