

Running Online Experiments

Session 2: crowdsourcing data
Justin Sulik justin.sulik@gmail.com
[@justinsulik](#)

Outcomes

- Have some idea of the strengths and weaknesses of online studies
- Know a few ways to manage attention
- Be aware of resources for managing participation
- Understand how to post a HIT on MTurk

Crowdsourcing

- Instead of paying people to come into the lab, just get people online to do it
- Various platforms provide participants:
 - Mechanical Turk
 - Crowdfunder
 - Prolific
 - ...

Pros

- Can pay people for short tasks (seconds, minutes, longer)
- Very quick data collection!
- Get a much bigger N
- Cost-effective
- Faster theory \leftrightarrow experiment cycle
- *Can* get wider participation

Cons

- You can't check up on what they're doing
- Non-naivete
- Occasional bad work (Though I haven't found it worse than the lab!)
- **Some** people try game the system
 - Lie about demographics (~3%)
 - Use bots
 - Answer very quickly

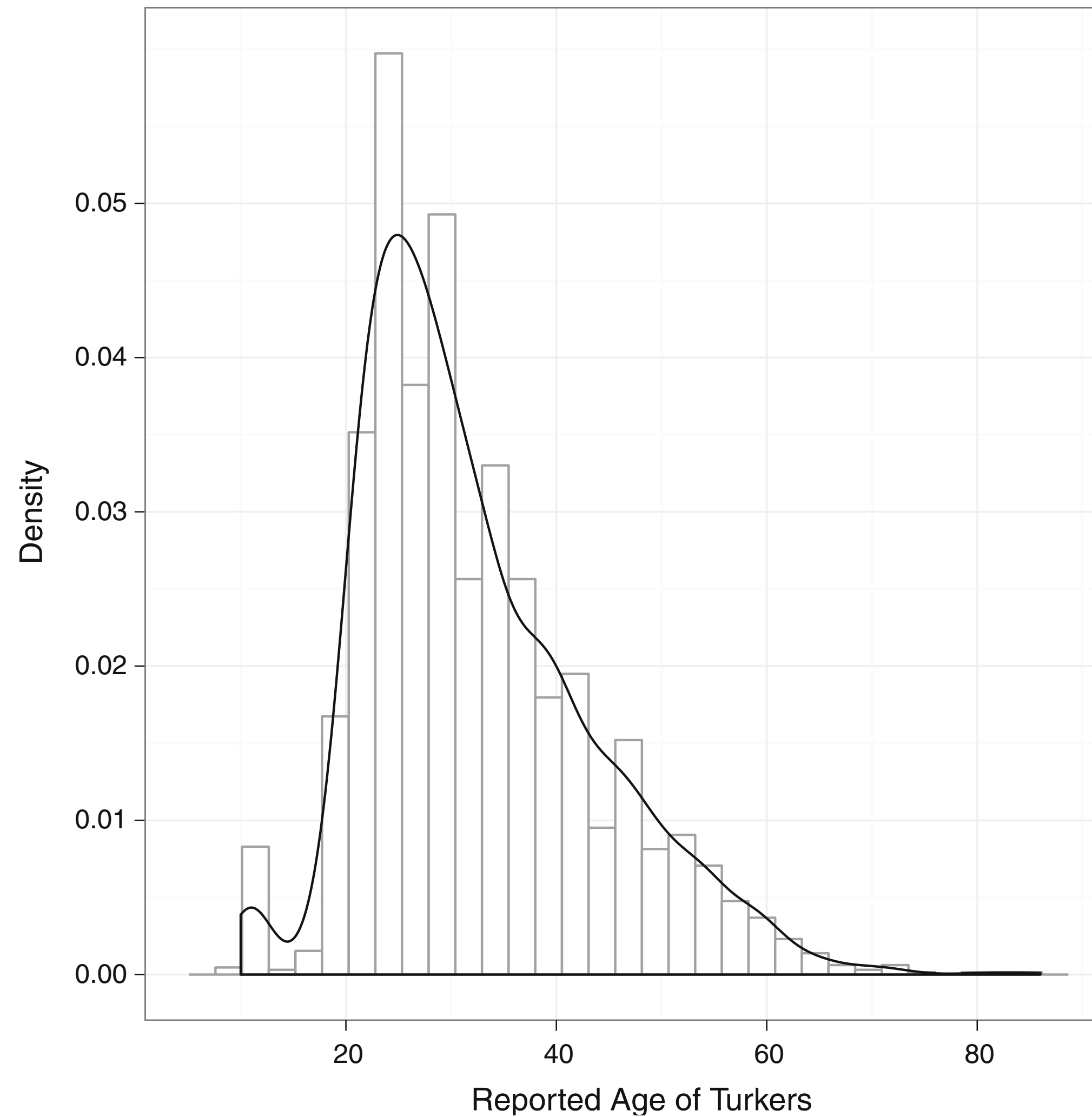
Crowdsourcing

- It works great, IF you manage
 - expectations
 - attention
 - participation
- Above all, communicate, be honest, and don't treat them like idiots

Who are they?

- Mostly American and Indian
 - You can specify particular countries
 - Work is **much** slower if you want something other than USA/India
- ~55% female, 45% male
- Modal education: bachelors
- Slightly more introverted than population avg.

Who are they?



Who are they?

- Earn ~\$30k/year
 - Including from other jobs!
 - Money is a big motivation, but not the only one
 - ~40% cite enjoyment as a major motivation

How good is their work?

- Good! Especially considering the price!
- Plenty of studies you can cite confirming this
- Compares well with lab performance
- Especially if you can get more responses & control for ESL
- The main challenge is adapting your mind-frame

How good is their work?

Behav Res

DOI 10.3758/s13428-011-0124-6

Conducting behavioral research on Amazon's Mechanical Turk

Winter Mason • Siddharth Suri

How good is their work?

OPEN  ACCESS Freely available online



Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research

Matthew J. C. Crump^{1*}, John V. McDonnell², Todd M. Gureckis²

How good is their work?

Journal of Behavioral Decision Making, J. Behav. Dec. Making (2012)

Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/bdm.1753

Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples

JOSEPH K. GOODMAN¹, CYNTHIA E. CRYDER^{1*} and AMAR CHEEMA²

Any issues?

- Their responses related to risk/money may be a bit atypical
- They do lots of similar experiments
 - (e.g. the CRT is tricky)
 - Make it different
 - Keep them on their toes
 - Check!
- Don't do long studies (I've never done > 15 mins)

Jargon

- HIT - human intelligence task
- Requester vs worker
- Turkopticon/turker nation
- Attention checks
- Rejection

Basics

- Payment \$1 for 10 mins - minimum!
 - Bear in mind they're not having to travel to the lab
 - Payment doesn't have a huge effect on quality of work
 - But it will affect your reputation
 - Be nice!
- Fees to Amazon: 20%
 - >10 workers/HIT = 40%
 - You can usually avoid this (coming up)

Treat it like a regular experiment

- Include proper consent forms
 - You don't have a signature, so make it impossible to continue unless they consent
 - Or end experiment if they click no

Attention check

- “While watching television, have you ever had a fatal heart attack?”
 - ~95%
- Word meanings
- Counter-intuitive instructions
- But bear in mind cost/time

Study 2

Research in decision making shows that people, when making decisions and answering questions, prefer not to pay attention and minimize their effort as much as possible. Some studies show that over 50% of people don't carefully read questions. If you are reading this question and have read all the other questions, please select the box marked 'other' and type 'Decision Making' in the box below. Do not select “predictions of your own behavior.” Thank you for participating and taking the time to read through the questions carefully!

What was this study about?

- A Predictions of your own behavior
- B Predictions of your friends' behavior
- C Political preferences
- D Other _____

Goodman, Cryder, Cheema (2012)

Rejects

- If they do badly or fail an attention check, you can reject their work
 - They don't get paid
 - It affects their reputation
 - If they grumble on forums, it can affect your reputation!
- I typically don't reject bad work
- But I do for attention checks IF I explain clearly what these will be like AND include a detailed description of what they did wrong when I reject their work

Good practice

- Keep it brief
- Keep instructions short
- Time everything
 - They could open the experiment, wander off for 10 mins
- Check understanding/attention
- Avoid google-able tasks
 - Or ask them not to
- Pilot/adjust/inform

Mechanical Turk is a marketplace for work.

We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

267,203 HITs available. [View them now.](#)

Make Money by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. [Find HITs now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



or [learn more about being a Worker](#)

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. [Get Started.](#)

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results



Categorization on Mechanical Turk

The Mechanical Turk **Categorization App** makes it simple to get fast, accurate results on your Categorization project!

- ✓ Quick and easy HIT design
- ✓ Pre-qualified workers
- ✓ Start receiving results in minutes

Create a Categorization Project



Work Distribution Made Easy

Mechanical Turk gives businesses and developers access to an on-demand, scalable workforce



Categorization on Mechanical Turk

The Mechanical Turk **Categorization App** makes it simple to get fast, accurate results on your Categorization Project!



Sentiment Rating Simplified

The Mechanical Turk **Sentiment App** makes it simple to collect and understand sentiment on your data!



- Can keep simpler tasks on Turk itself
 - e.g. \$0.02 to rate how similar two pictures are
- OR provide link to experiment and give them randomly generated completion code at end to enter on turk
 - In which case, make sure you leave somewhere to enter it!

[Edit Project](#)

Use the HTML editor below to design the layout of your HIT. This layout is common for all of the HITs created with this project. You can define variables for data that will vary from HIT to HIT ([Learn more](#)).

1 Enter Properties

2 Design Layout

③ Preview and Finish

Project Name: crossModality_affect

This name is not displayed to Workers.

Frame Height

450

Height in pixels of the frame your HIT will be displayed in to Workers. Adjust the height appropriately to minimize scrolling for Workers.

Format


Font

U

I

B

A →

$$\underline{I}_x$$


三

1=
2=

—

65

Source

We are conducting an experiment to see how people match stimuli such as colors, shapes and sounds. The study takes about 10 minutes.

You will not be able to take part if the display area of your browser is less than 900x640px, so you will not be able to use a smartphone or tablet. The display area does not include tool bars and similar, so even if your screen is large enough, it might be that the display area is too small.

You will also need to be able to hear audio stimuli (preferably with headphones). This will be tested during the practice rounds, and if you answer those questions incorrectly, your work will be rejected, so don't accept this HIT if you are unable to play/hear audio.

Make sure to leave this window open as you complete the survey. When you are finished, you will return to this page to paste the code into the box.

[Click here to go to the study](#)

Completion code:

Any comments?

- turkPage.txt (catsperiment) has the code (basic but functional)
- You might notice $\${condition}$ in the script
- If you include $\${x}$ then it will ask you to upload a csv with all the variables you need to use for each HIT
 - x as column
 - In that case, the number of HITS on the setup page mean **per line**

- <http://www.myUni.edu/~myname/catsperiment.html?condition=1>
- [http://www.myUni.edu/~myname/catsperiment.html?condition=\\${condition}](http://www.myUni.edu/~myname/catsperiment.html?condition=${condition})
- In experiment script:

```
var condition = jsPsych.data.getURLVariable('condition');
```
- Let's try now!

Managing participation

- You presumably don't want people to take part multiple times
- If you've got it all in one HIT, you don't have to worry
- If you spread it across multiple HITS (or multiple studies) you might want to prevent someone responding more than once

Managing participation

- On turk: assign qualification
 - I have a python script that goes through the data csvs, extracts their workerId, and creates a list of people I won't use again
 - Google “assign qualifications”
 - Be aware of encoding issues!
 - There are automatic ways to handle this without downloading/uploading csvs, but we can't get into those now
- Or use turkgate/turkprime

When?

- Pretty much any time
- I'm working on a study that checks different times/days for performance
- I usually launch my studies around 9am EST

Managing participation

- Don't ask for personal info
- Don't waste their time
- Be creative about engagement
- Respond to emails!

Managing participation

- Let's say you have an experiment that requires two people to play a game
- (and you've set up the script to handle it)
- You need 2 turkers to have the window open for it to begin
- How do you think you'll manage their attention/engagement?

Managing noise

- Let's say you suspect that people are pressing buttons randomly. How will you manage their behavior/your data?
- Imagine that some of the text answers don't sound like good English. What would you do better the next time?
- Let's say that someone thought your instructions were terrible and they failed an attention check. How would you deal with their fury?

Hi,

With respect I could hear the dog/cat sounds just fine (for the record the top was cat and bottom was dog for the attention check)...but your instructions were severely unclear on how to submit the actual answers. They said to click on areas, yes, but clicking them did nothing visual to help, although clicking on other areas removed some black dots...I couldn't even figure out how to submit it correctly.

If you're going to make an attention check..please make sure the instructions are clear and not delegate it to a 0.25 bonus.

I'm returning the HIT because there is no point in taking a rejection with a 0.25 bonus for unclear instructions. Please...be more clear and at least give a visual aid to show that this dot has 'cat' or 'dog' selected.

If this is an issue in Firefox (the most recent version) and not your instructions I hope you will address it.

Hi

I'm sorry you found the task confusing. However, we've run this hundreds of times and have a failure rate of about 2%, which I think is quite normal for an MTurk experiment. If the instructions genuinely were unclear, the failure rate would not be that low.

Your reference to "some black dots" makes me think you might have missed the part of the instructions that explains that the black dots are how you make a choice, and how to interpret those dots to see which choice had been selected. Perhaps you accidentally hit space twice quickly, and thus skipped a page - I'm sure it was an innocent mistake. But we are reluctant to block people's ability to read through instructions quickly, since that would frustrate way more people.

I understand your reluctance to take a rejection, but sadly that means I can't bonus you for the time spent reading the instructions.

Regards
Justin

Hi Justin,

Sorry for the tone. I was having a frustrating day.

I did read the instructions, but maybe that just didn't register as much as it should have with the dots (at the time). If so, that's my fault and not yours. I can't remember the specific details anymore, but I believe you completely about the failure rate. I'd say that's pretty normal.

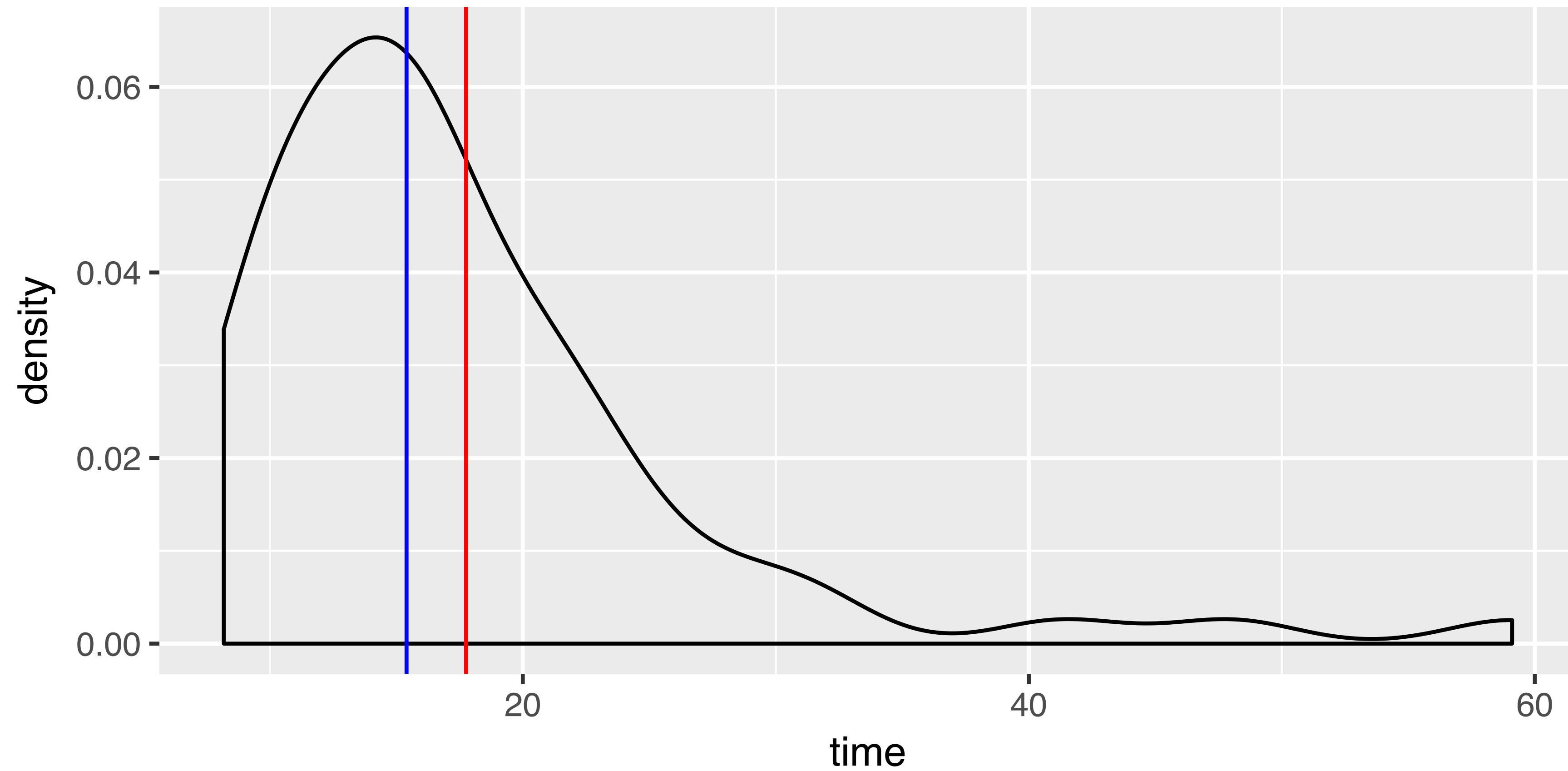
Again, I shouldn't have written my response that way and I apologize. I want requesters to stay responsive, so I feel it's important you hear that instead of a giant rant that was unfair of me.

I wasn't expecting a bonus, and I returned it so no harm done. If I see a study from you in the future I'll be more careful.

Thanks, and I appreciate the response!

Timing

- OMG IT TOOK ME AGES GIVE ME MONEY



Summary

- Use it - it's great resource
- If you're nervous (will my script work, will the `${}` variables get in the right place, will people mess it up),
 - Try it out with 1 HIT
 - At the worse, you waste \$0.10
 - Then add more things
- Test your experiments on your friends first
 - Get an idea of time
 - Make sure there are no bugs

Summary

- Be clear, be communicative, be fair
- But don't just assume they'll follow instructions/only continue if they're English speakers, etc.