Online Experiments for Language Scientists

Lecture 2: Crowdsourcing

Kenny Smith

kenny.smith@ed.ac.uk

What you will have read for today
(from https://kennysmithed.github.io/oels2024/oels_reading_wk2.html)

Reading tasks for this week

There are several things to read/look at this week, to give you a feel for some of the issues around online vs lab data collection, demographics of online populations, what online experiments look like, and what a crowdsourcing site looks like from the participant's perspective.

Read:

- The wikipedia page explaining what MTurk is.
- Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing Samples in Cognitive Science. *Trends in Cognitive Sciences, 21*, 736-748.

Also read *at least one of*:

- Monroe, R. et al. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130.
- Pavlick, E. et al. (2014). The Language Demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics, 2*, 79-92.
- A blogpost listing some downsides of MTurk, although note that this is written by people at Prolific, who are an MTurk competitor! Doesn't mean some of the points aren't valid though.
- A 2018 article in The Atlantic on worker exploitation on MTurk, although remember that we at least control what we pay workers and how we treat them.

It might also help to have a look at a video of what the MTurk site looks like from a worker's perspective - NB this video dates from December 2020. I'll show you around Prolific in class.

Crowdsourcing

Once you have an experiment that runs in a browser, you can get participants from anywhere, including crowdsourcing sites

- Websites with populations of “workers” who will do online tasks for money

MTurk and Prolific

Amazon Mechanical Turk

https://www.mturk.com

- Designed for crowdsourcing **anything**
- Very light touch
- More US-based participants?
- Interface is pretty horrible (particularly for experimenter) but has a powerful API for code-based payment etc
- More chaotic, worse data (or more need to restrict participation to established workers)?

Prolific (formerly “Prolific Academic”)

https://www.prolific.co

or https://www.prolific.com now

- Designed for scientific data collection
- Heavier vetting of participants
- More UK/EU participants?
- Nicer web interface, recently added API
- Maybe better-behaved participants

A look around Prolific

- From a participant perspective
- From an experimenter perspective

Pros and cons of crowdsourcing experimental data

Pros

- **Large samples, fast**
- Access different populations
- + for replicability

Cons

- Expensive (**not** cheap)
- Lack of control
- Encourages dumb experiments?
- - for replicability

Pro: large samples, fast

MTurk and Prolific both have large active populations of workers/participants (100,000s of registered people)

- Although not everyone is active all the time
- Estimating Mturk population size is complicated (see e.g. Difallah et al., 2018)
- Prolific gives you an estimate of available and active population size

In practice, you can recruit **100s/1000s of participants in days**.

Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining.* https://doi.org/10.1145/3159652.3159661

Pro: access different populations

Typical lab-based studies will sample from university student population

- Mostly undergraduates
- Mostly young
- All highly educated
- Here, mainly native English speakers (obviously varies between unis)

If you want to access a different population, crowdsourcing might let you do that

From Pavlick et al. (2014)

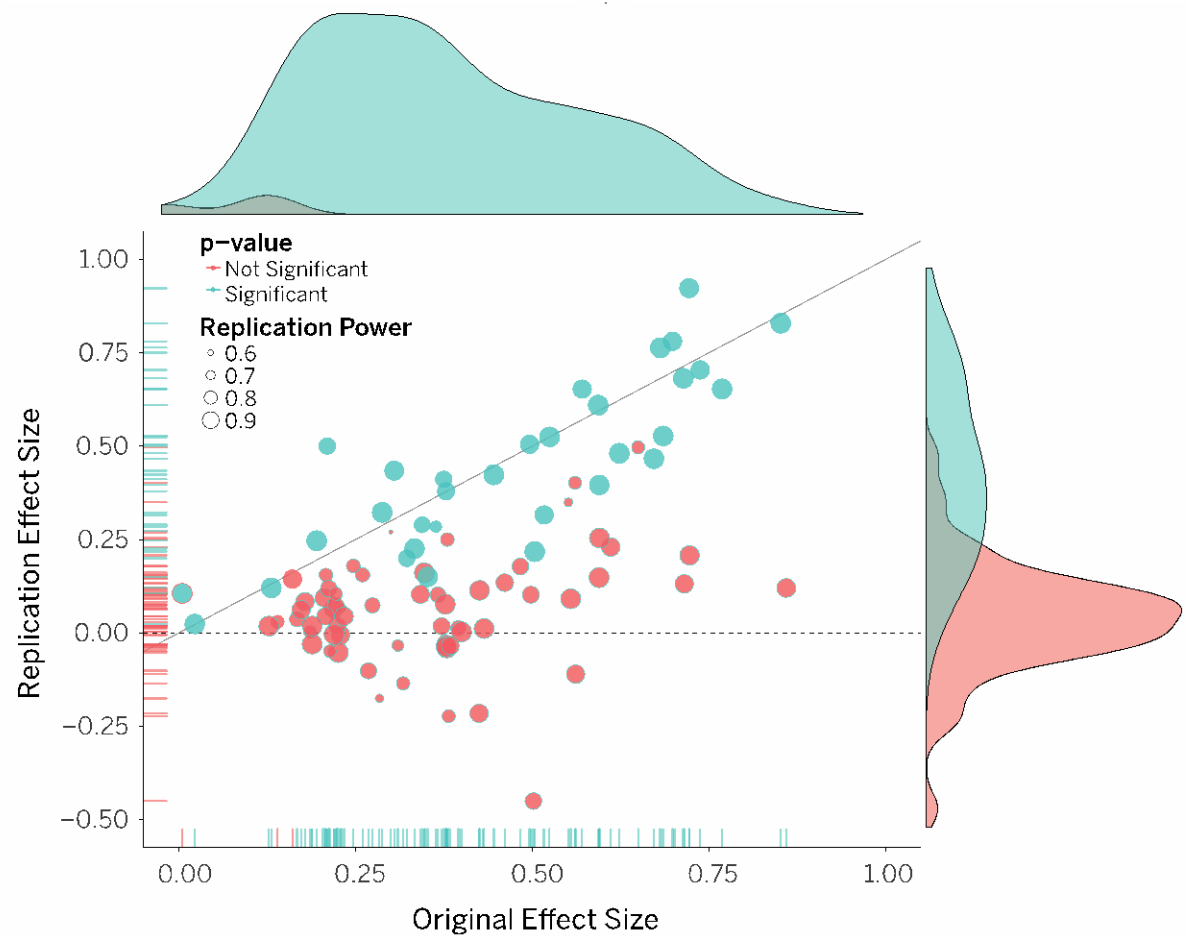
workers	quality	speed	
many	high	fast	Dutch, French, German, Gujarati, Italian, Kannada, Malayalam, Portuguese, Romanian, Serbian, Spanish, Tagalog, Telugu
		slow	Arabic, Hebrew, Irish, Punjabi, Swedish, Turkish
	low or medium	fast	Hindi, Marathi, Tamil, Urdu
		slow	Bengali, Bishnupriya Manipuri, Cebuano, Chinese, Nepali, Newar, Polish, Russian, Sindhi, Tibetan
few	high	fast	Bosnia, Croatian, Macedonian, Malay, Serbo-Croatian
		slow	Afrikaans, Albanian, Aragonese, Asturian, Basque, Belarusian, Bulgarian, Central Bicolano, Czech, Danish, Finnish, Galacian, Greek, Haitian, Hungarian, Icelandic, Ilokano, Indonesian, Japanese, Javanese, Kapampangan, Kazakh, Korean, Lithuanian, Low Saxon, Malagasy, Norwegian (Bokmal), Sicilian, Slovak, Slovenian, Thai, UKranian, Uzbek, Waray-Waray, West Frisian, Yoruba
	low or medium	fast	–
		slow	Amharic, Armenian, Azerbaijani, Breton, Catalan, Georgian, Latvian, Luxembourgish, Neapolitian, Norwegian (Nynorsk), Pashto, Piedmontese, Somali, Sudanese, Swahili, Tatar, Vietnamese, Walloon, Welsh
none	low or medium	slow	Esperanto, Ido, Kurdish, Persian, Quechua, Wolof, Zazaki

Pro: + for replicability

If you see a result in a scientific paper, can you assume that the effect they report is real and not just, e.g., a statistical fluke?

One way to check: replication

- Take someone else's experiment, replicate it, check you get the same result



From Open Science Collaboration, 2015, *Science*

Pro: + for replicability

If you see a result in a scientific paper, can you assume that the effect they report is real and not just, e.g., a statistical fluke?

One way to check: replication

- Take someone else's experiment, replicate it, check you get the same result

Multiple potential advantages for online data collection

- Because collecting a large sample is easy, small-sample experiments (which are more prone to statistical flukes) can be avoided
- Because collecting data online is fast and easy, it makes it easier to replicate experiments (including your own!)
- Because populations are shared, makes it easy to replicate closely (avoiding e.g. "ah it's because your population is different" responses to non-replication)

Con: expensive (not cheap)

Mturk does not set minimum pay rates

Prolific does, but they are low (£6/hour)

Current rates

These rates are for the National Living Wage (for those aged 21 and over) and the National Minimum Wage (for those of at least school leaving age). The rates change on 1 April every year.

	21 and over	18 to 20	Under 18	Apprentice
April 2024	£11.44	£8.60	£6.40	£6.40

https://www.gov.uk/national-minimum-wage-rates

Con: expensive (**not cheap**)

Mturk does not set minimum pay rates

Prolific does, but they are low (£6/hour)

But we should not be paying at those rates

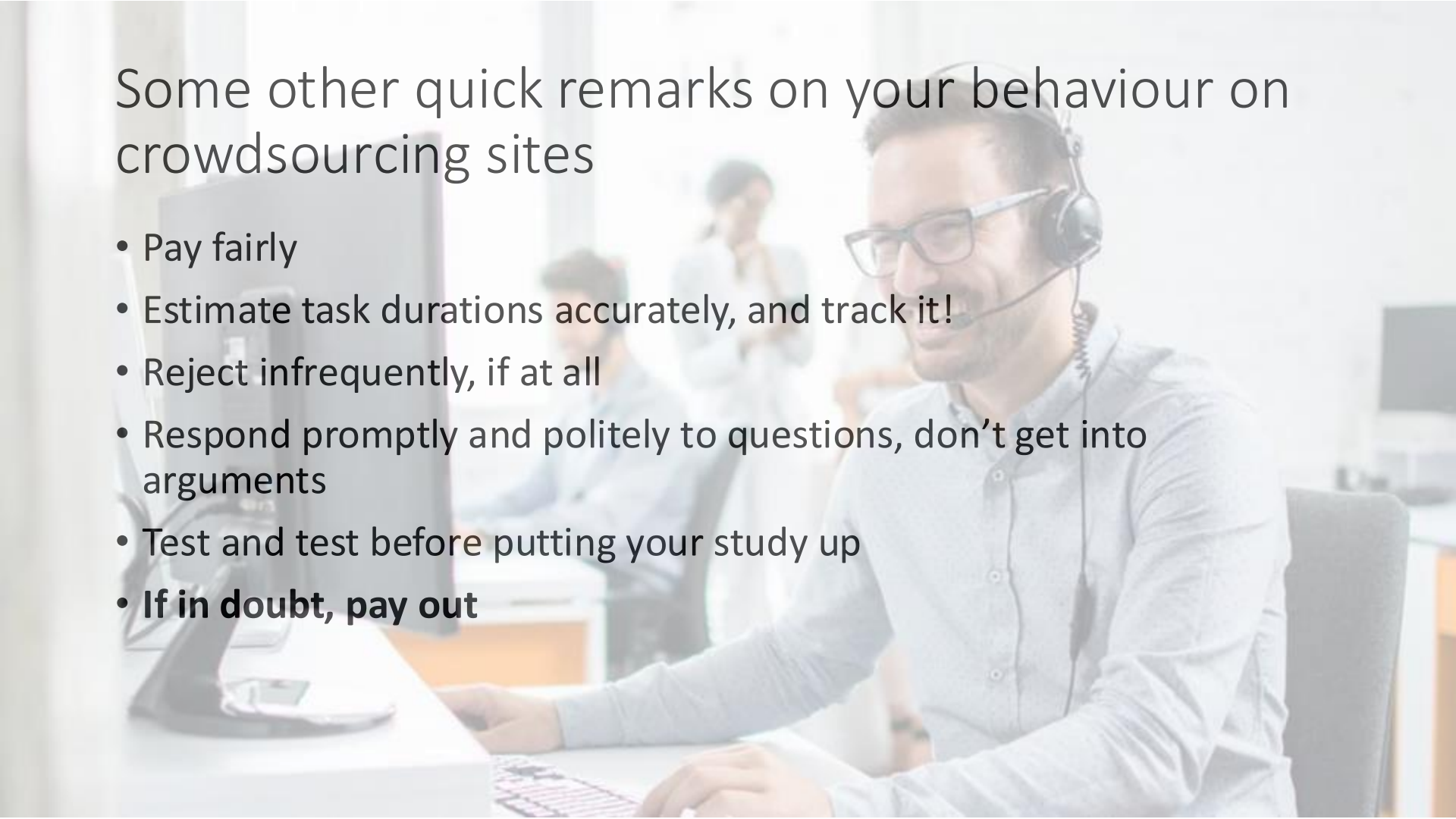
- It's unethical
- It's exploitative

Additionally

- Mturk and Prolific charge fees: 20-40% on top of what goes to participant
- Plus sample sizes tend to be bigger (because data quality can be lower or just because you can)

Some other quick remarks on your behaviour on crowdsourcing sites

- Pay fairly
- Estimate task durations accurately, and track it!
- Reject infrequently, if at all
- Respond promptly and politely to questions, don't get into arguments
- Test and test before putting your study up
- **If in doubt, pay out**



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.

Con: encourages dumb experiments (?)

No hard constraints, but because of the lack of control, stuff that works best involves constrained and low-effort responses

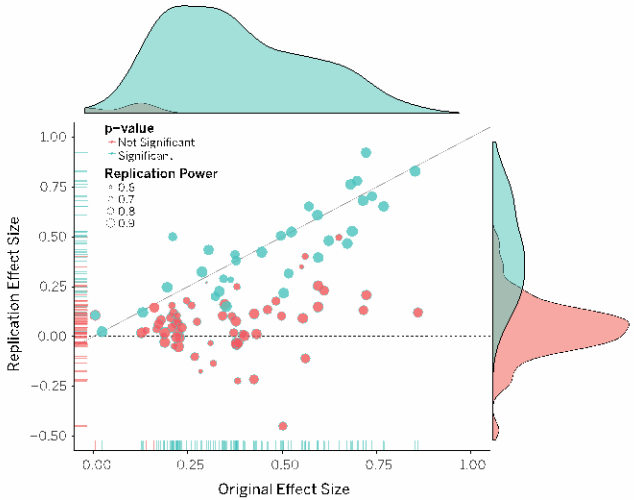
- One-off decisions (i.e. not involving complex integration of info)
- Few restricted choices per trial (not e.g. open-ended typing)
- Short experiments (a few minutes rather than an hour)

Can you investigate the questions you want using these sorts of methods?

Con: - for replicability

If you see a result in a scientific paper, can you assume that the effect they report is real and not just, e.g., a statistical fluke?

Potential risk of online data collection: Because collecting data online is fast and easy, it makes it easier to run lots of experiments, and just publish the ones that “work” (cf. “the file drawer problem”)



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. from this week’s set readings, Monroe et al., 2010)

Time for Q&A/discussion on this week’s readings?

Next up

Wednesday, 9am: lab

- More basics of jsPsych and javascript
- Come prepared!

Next Monday: grammaticality judgments

- Do the reading beforehand!