Online Experiments for Language Scientists, UoB

Lecture 1a: Intro + co-speech gesture

Kenny Smith

kenny.smith@ed.ac.uk

Three components of running an online experiment

Building an experiment that will run in a web browser

- We'll be using javascript and jsPsych
- Also useful for running experiments in-person!

Making it openly available online

- Uni or commercial servers

Connecting with experiment participants

- E.g. through **crowdsourcing websites**

A look at some simple experiments

Javascript and jsPsych

Javascript: a programming language that runs in web browsers

jsPsych: a library that makes it easy to build experiments (https://www.jspsych.org)

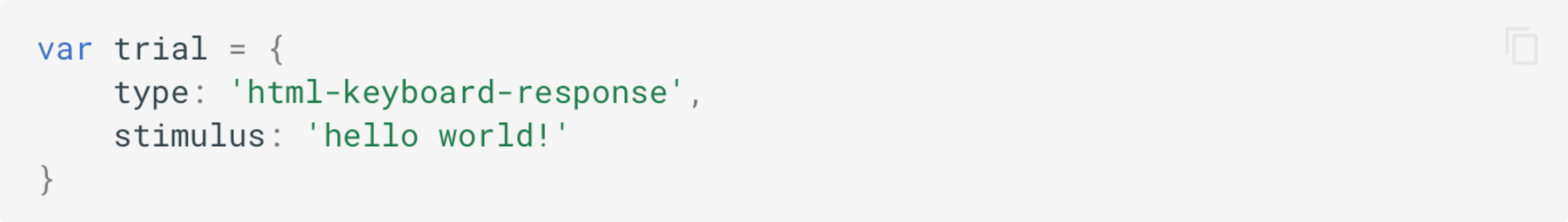
de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*, 1-12. doi:10.3758/s13428-014-0458-y.

Josh de Leeuw
Vassar College

2

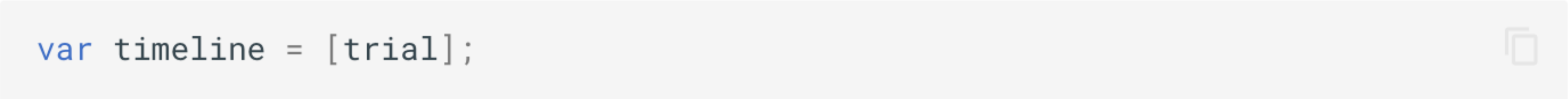
Plugins and timelines

Plugins: basic building blocks



```
var trial = {
    type: 'html-keyboard-response',
    stimulus: 'hello world!'
}
```

Timeline: a sequence of those building blocks



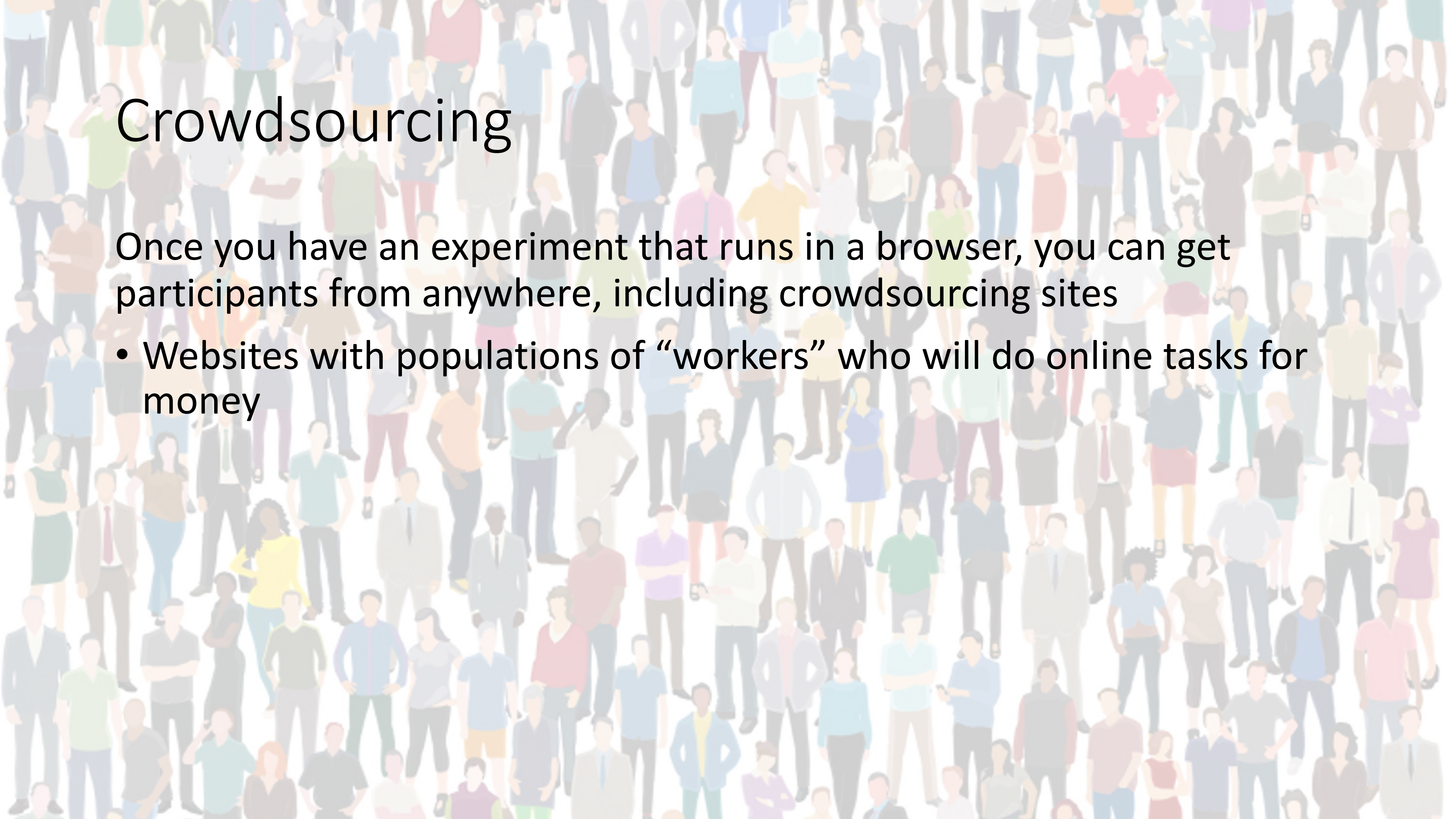
```
var timeline = [trial];
```

A wide range of plugins available

See https://www.jspsych.org/6.3/plugins/list-of-plugins/

Building an experiment involves

- Knowing how to use plugins
- Figuring out how to piece them together to make the experiment you want
- Some tiny bits of html and javascript to connect the plugins and make them do what you want
- (Occasionally, and optionally, making your own plugin)



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

- Designed for scientific data collection
- Heavier vetting of participants
- More UK/EU participants?
- Nicer web interface, but no 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

- Not face-to-face
- Large samples, fast
- Access different populations
- + for replicability

Cons

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

Pro: not face-to-face

Remember
FACTS
for a safer Scotland
Phase 3
2M
Face coverings
Avoid crowded places
Clean your hands regularly
Two metre distance
Self isolate and book a test if you have symptoms

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**.

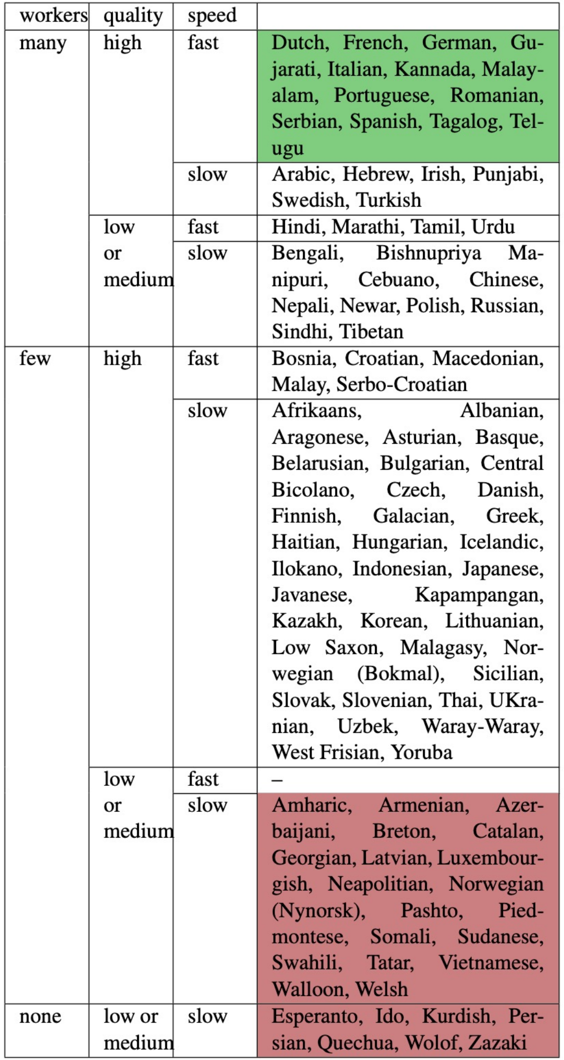
Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining.* https://doi.org/10.1145/3159652.3159661

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



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, Neapolitan, Norwegian (Nynorsk), Pashto, Piedmontese, Somali, Sudanese, Swahili, Tatar, Vietnamese, Walloon, Welsh
none	low or medium	slow	Esperanto, Ido, Kurdish, Persian, Quechua, Wolof, Zazaki

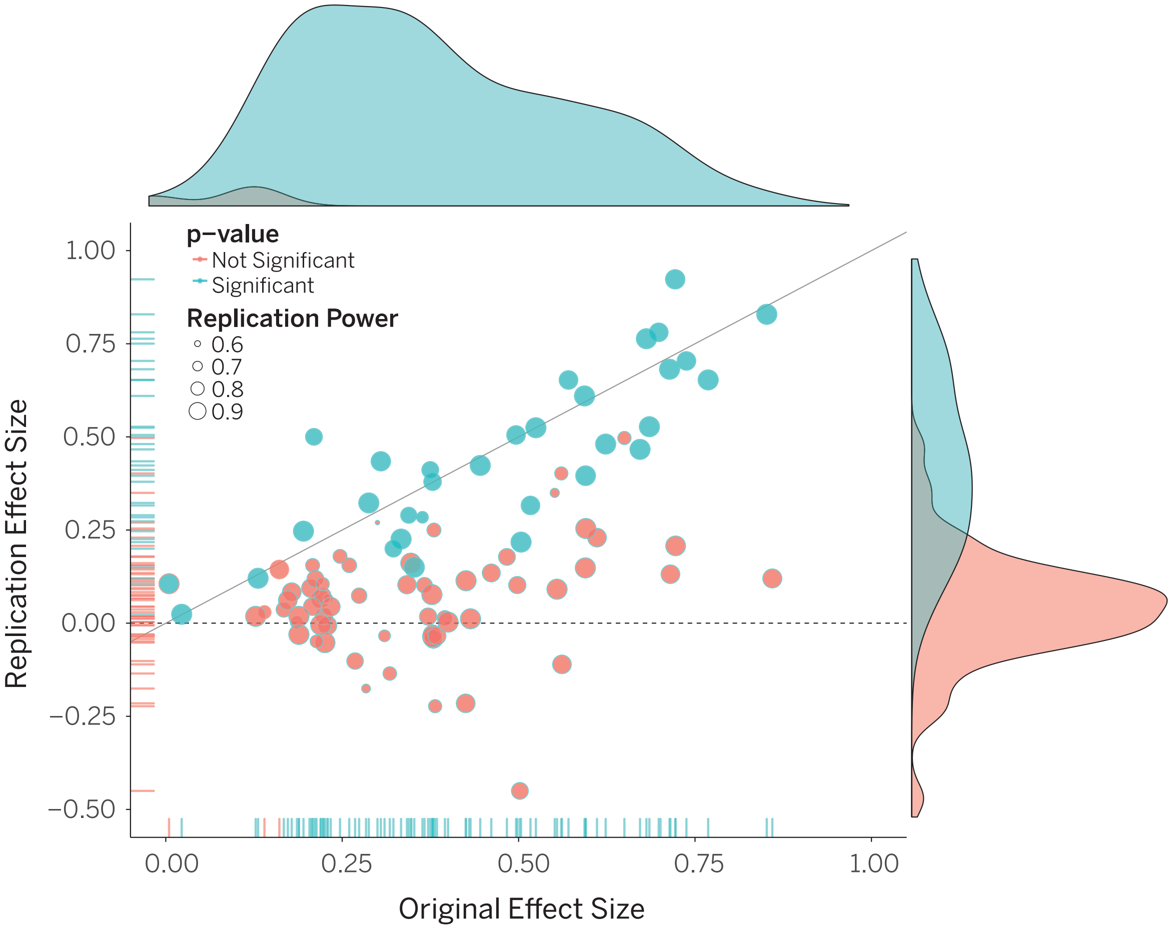
From Pavlick et al. (2014)

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

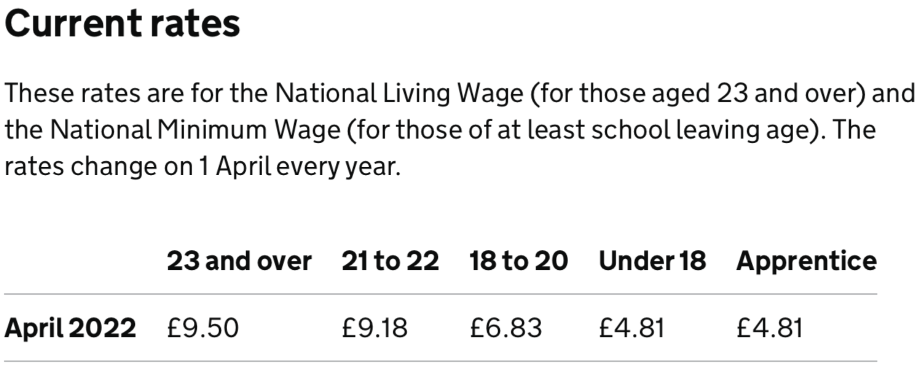
Multiple potential advantages for online data collection

- 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 23 and over) and the National Minimum Wage (for those of at least school leaving age). The rates change on 1 April every year.

	23 and over	21 to 22	18 to 20	Under 18	Apprentice
April 2022	£9.50	£9.18	£6.83	£4.81	£4.81

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)

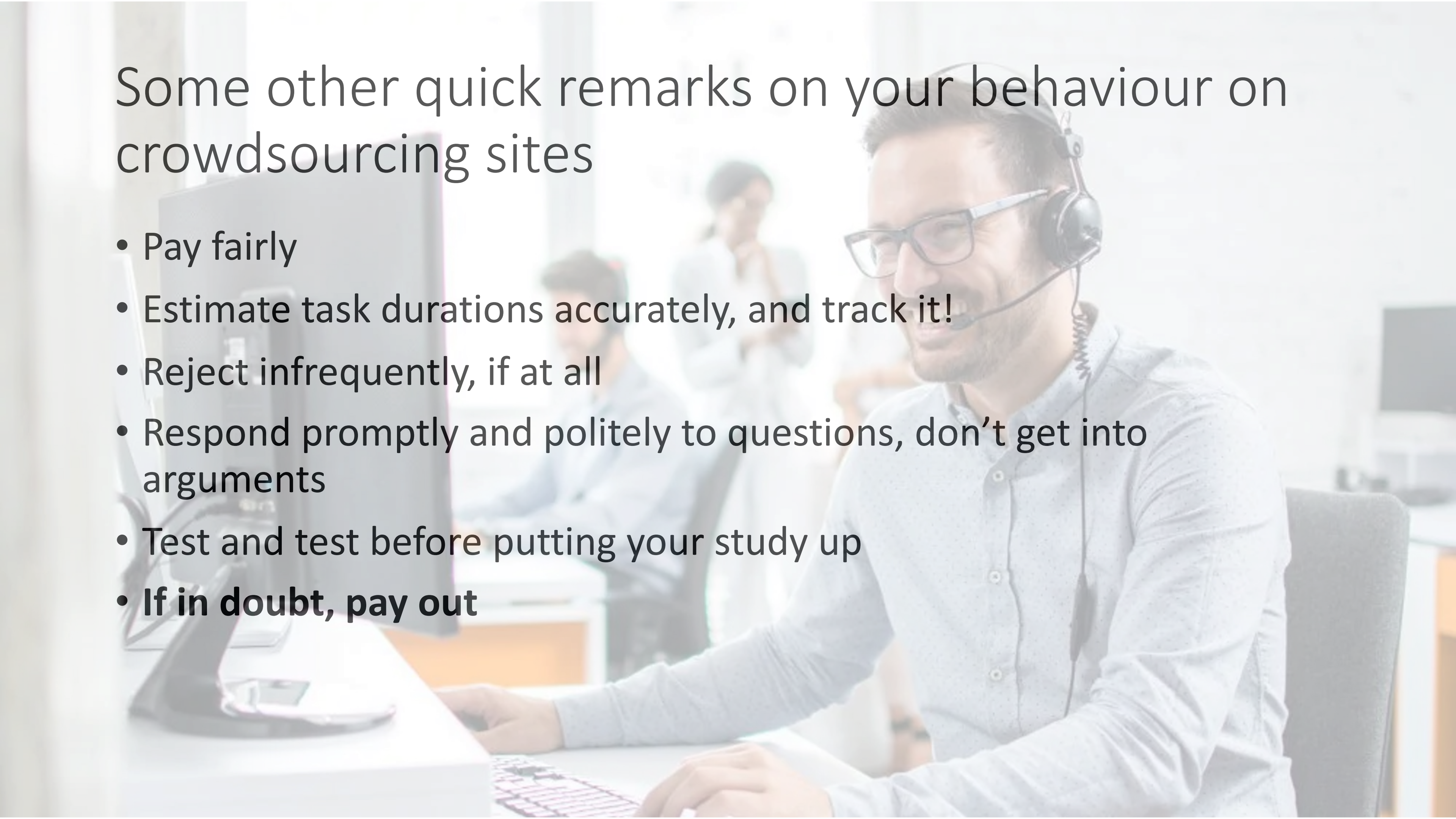
But we should not be paying at those rates

- 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

Player: PSSH. Good luck hitting me. My AC is incredible.

DM:

- 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

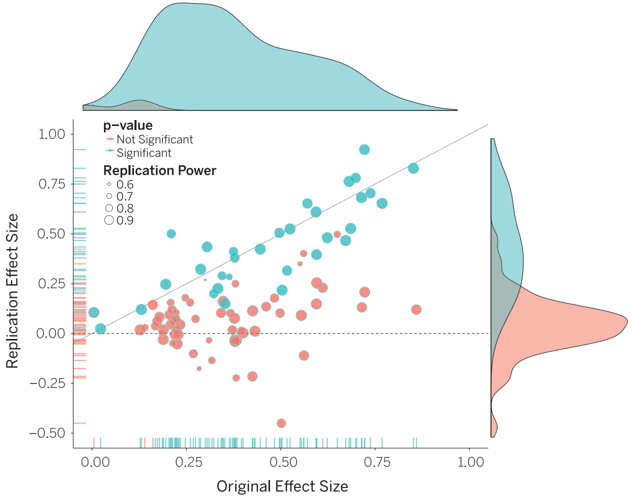
- One-off decisions (i.e. not involving complex integration of info)
- Few restricted choices per trial (not e.g. open-ended typing)
- Short experiments (a few minutes rather than an hour)

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 several papers making direct comparisons

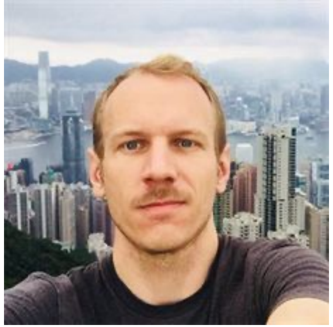
Our first experiment: co-speech gesture

Winter & Duffy (2020)

Winter, B., & Duffy, S. E. (2020). Can Co-Speech Gestures Alone Carry the Mental Time Line? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46,* 1768-1781.

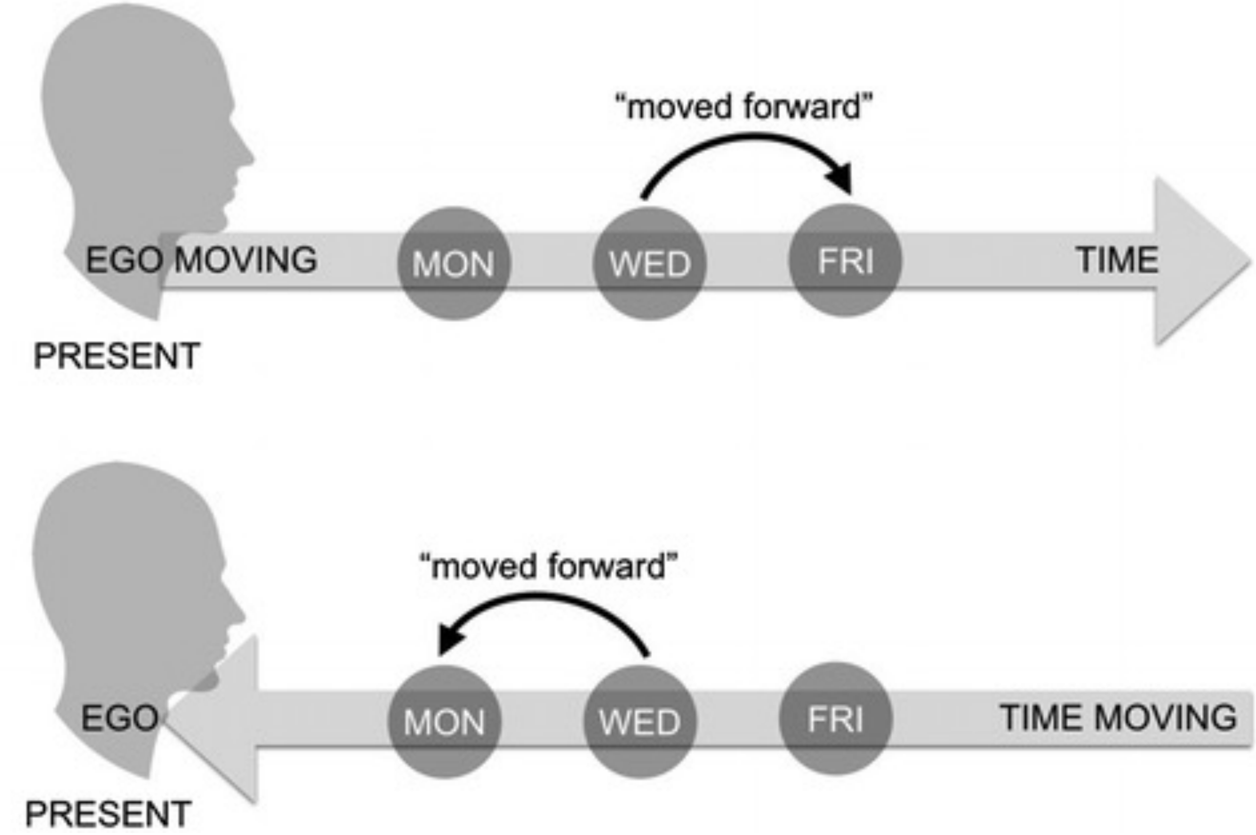
Interaction between speech and gesture in interpretation

- Is interpretation of ambiguous "next Wednesday's meeting has been moved forward/backward 2 days" more influenced by adverb or gesture?



Bodo Winter
University of Birmingham

Sarah Duffy
Northumbria University



From Stocker, K., & Hartmann, M. (2019). "Next Wednesday's Meeting has been Moved Forward Two Days". *Swiss Journal of Psychology*, *78*, 61-67

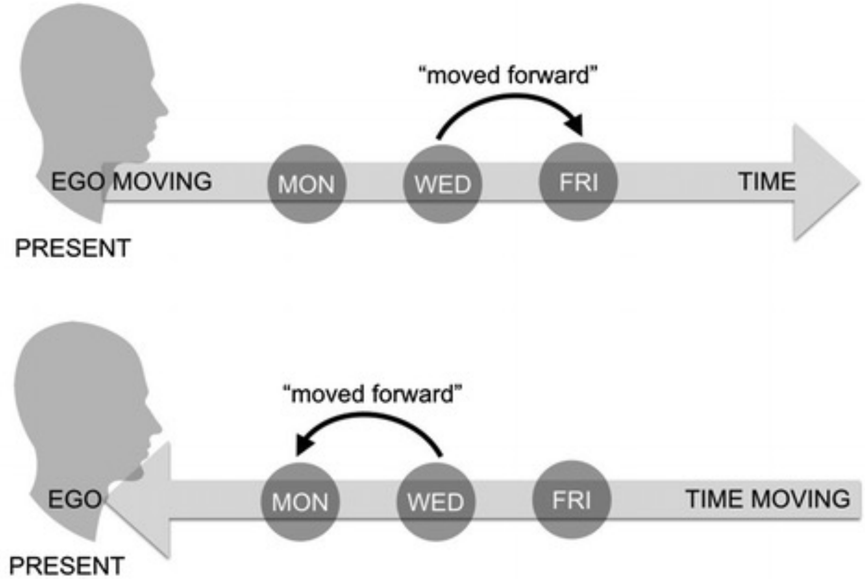
Perspective can be altered by gesture or words

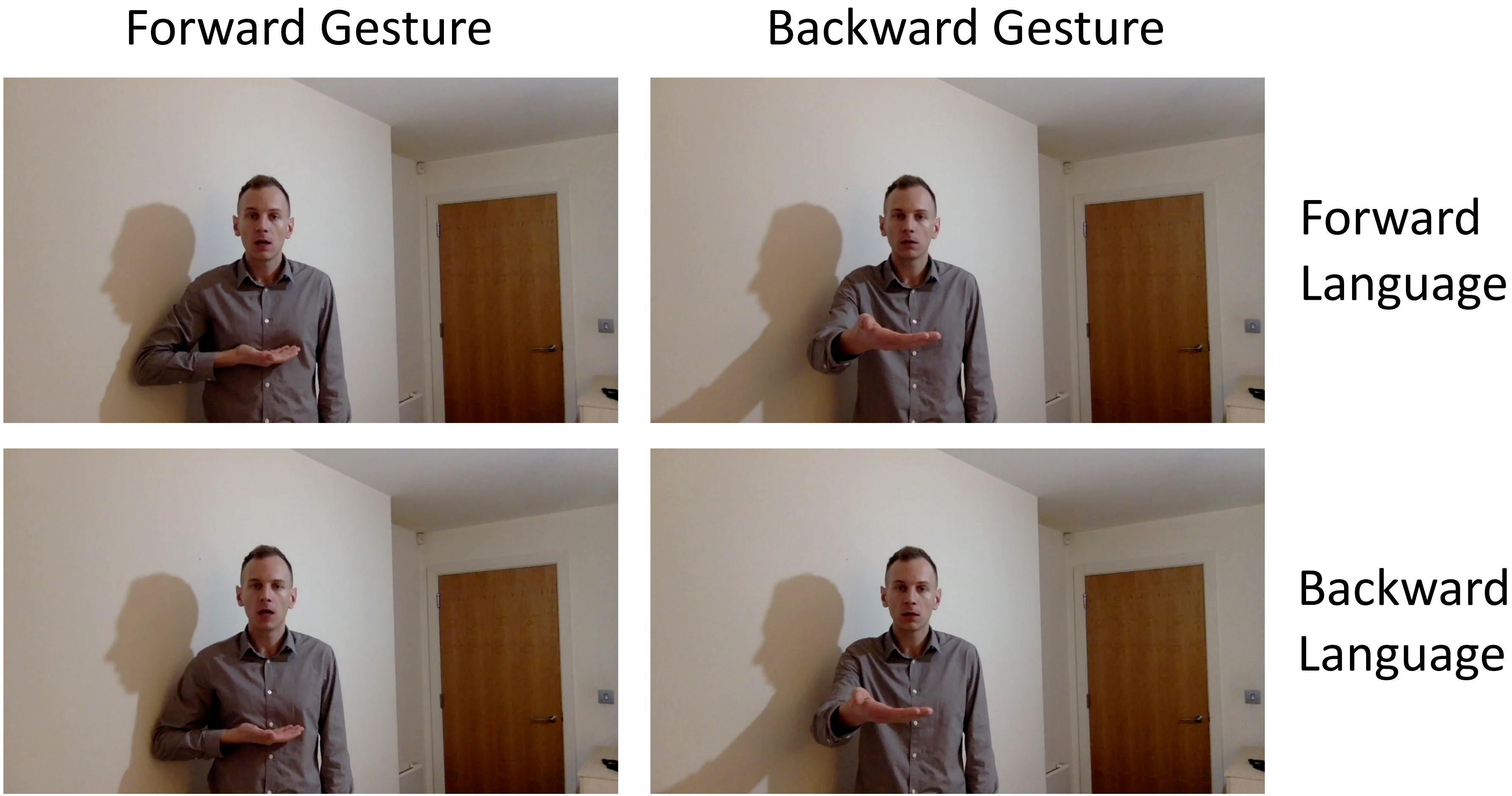
Gesture (Lewis & Stickles, 2017)

- Away from speaker -> more Friday
- Towards speaker -> more Monday

Adverb (Feist & Duffy, 2015)

- ”forward” -> more Monday
- “backward” -> more Friday

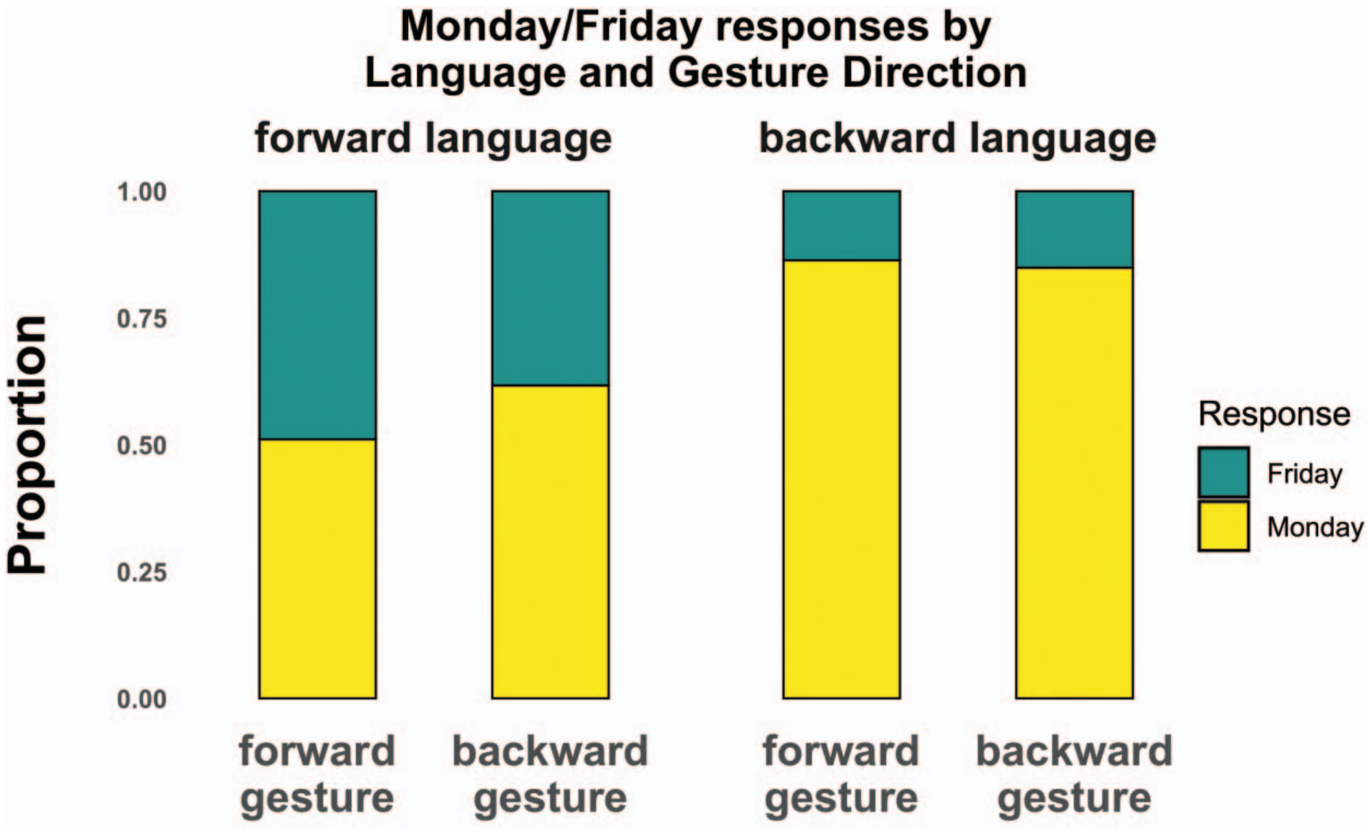


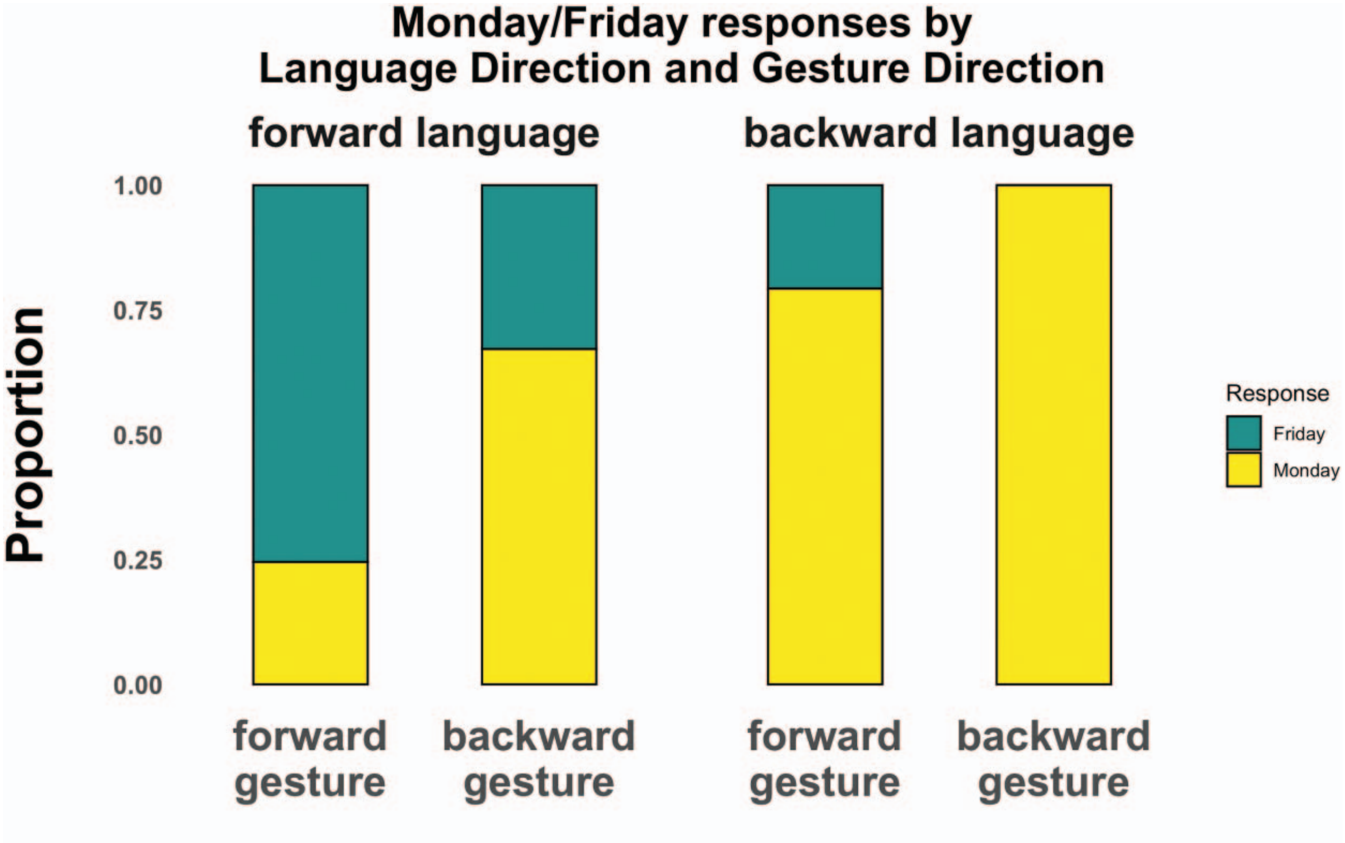

Forward Gesture
Backward Gesture
Forward Language
Backward Language

Sample size etc

- Watch video, give free text response
- US-based MTurk workers with 92%+ approval
- N=191 after exclusions (12 excluded for failing arithmetic attention check Q, 36 for non Monday/Friday responses)
- $0.30 (duration likely a few minutes?)

Demo using our code

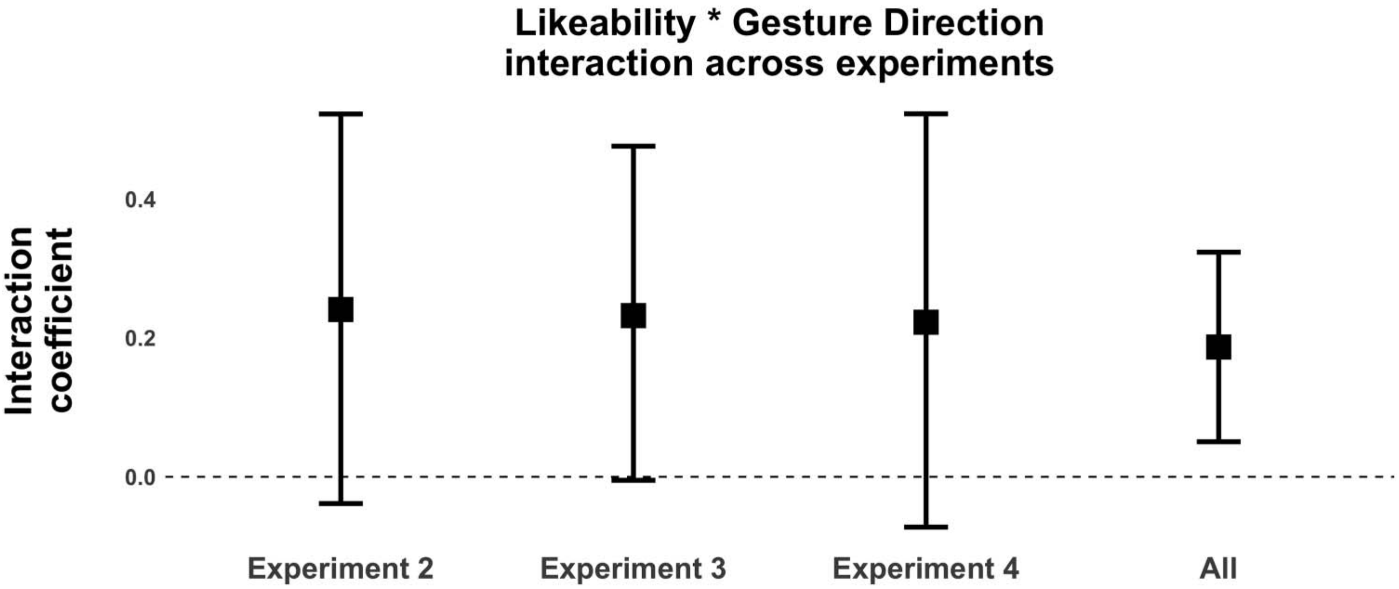

Monday/Friday responses by
Language and Gesture Direction
forward language
backward language
Proportion
1.00
0.75
0.50
0.25
0.00
forward gesture
backward gesture
forward gesture
backward gesture
Response
Friday
Monday


Monday/Friday responses by
Language Direction and Gesture Direction
forward language
backward language
Proportion
1.00
0.75
0.50
0.25
0.00
forward gesture
backward gesture
forward gesture
backward gesture
Response
Friday
Monday

An additional explanation is suggested by the participants' responses to the debriefing question. Some had very positive impressions of the speaker saying such things as “He was very articulate” or “He seemed like a nice guy.” Others noted that the speaker seemed strange or foreign, such as “He isn't native to the United States, and uses odd hand gestures,” “Is he from a different planet?,” and “He seems like he wants to be a mentalist.” We thought that perceived likability of the speaker may be a moderating factor in this experiment since the directional effect of the gesture depends on whether one is willing to assume the speaker's perspective or not. Thus, Experiment 2 included two socially relevant scales to investigate this phenomenon.

...

To explore social factors in influencing perspective taking, we added four items from Reysen's likability scale (Reysen, 2005), asking whether participants thought that the “person in the video” is “warm,” “approachable,” “friendly,” and “likeable.”


Likeability * Gesture Direction
interaction across experiments
Interaction coefficient
0.4
0.2
0.0
Experiment 2
Experiment 3
Experiment 4
All

Demo using our code