Perception, Memory, and Coordination

Alexandra Paxton
paxton.alexandra@gmail.com
Institute of Cognitive and Brain Sciences
Berkeley Institute for Data Science
University of California, Berkeley

Thomas J. H. Morgan
thomas.j.h.morgan@asu.edu
School of Human Evolution and Social Change
Arizona State University

Jordan W. Suchow
suchow@berkeley.edu
Social Science Matrix
University of California, Berkeley

Thomas L. Griffiths
tom_griffiths@berkeley.edu
Department of Psychology
University of California, Berkeley

Abstract

With cognitive scientists' increasing interest in moving outside of the lab, recent advances in crowdsourcing platforms can help strike a balance between the tight experimental control of lab designs and the affordances of web-based experiments to reach beyond traditional undergraduate subject pools. By taking advantage of new tools, scientists interested in social cognition and behavior can create new designs and adapt traditional ones to deliver experiments at scale. Dallinger is one such tool, providing researchers with an open-source experiment platform that provides end-to-end automation of the experiment pipeline, from participant recruitment and consent to data de-identification and participant compensation. Here we demonstrate how Dallinger can be used to run complex experimental studies of interactive human social behavior, as a demonstration of its potential to study social cognition and behavior using designs drawn from across cognitive science.

Keywords: interpersonal interaction; human communication; crowdsourcing; Dallinger

Introduction

Today, cognitive science is more interested in expanding its horizons than ever. With growing excitement around the value of big data and naturally occurring datasets (Goldstone & Lupyan, 2016), data like social media activity and video game logs hold the promise of capturing behavior in the wild and providing a testing ground for key scientific theories (Paxton & Griffiths, 2017). While these data can provide a window into observational data about human behavior at a massive scale, technological advances are quickly expanding to accommodate new *experimental* paradigms as well.

Crowdsourcing platforms like Amazon Mechanical Turk (`http://www.mturk.com`) have been a source for data collection for the better part of relatively simple but robust experimental paradigms—like surveys (Buhrmester, Kwang, & Gosling, 2011) and mouse-tracking (Freeman, Dale, & Farmer, 2011)—for the better part of the last decade. The earliest work in this domain required researchers to use established survey creation tools or to program bespoke experiments, but over the past several years, cognitive scientists have worked to create solutions to support a range of experiments (e.g., Gureckis et al., 2016). As the community around online psychology experiments has grown, it has done so with the intent to broaden its reach (especially to researchers with less programming experience) and to continuing to provide more powerful experimental tools.

We believe that Dallinger can provide researchers interested in social behavior the opportunity to expand their experimental capabilities beyond the lab while not compromising on the richness and complexity of true interactive contexts.

Dallinger

Jordan and Tom — maybe y'all can add some text and relevant citations here?

(should probably include some citations about Turk populations, too)

Social Behavior: Interpersonal Coordination

The Present Study

Method

All research activities were completed in compliance with oversight from Committee for the Protection of Human Subjects at the University of California, Berkeley.

Participants

Participants (n = 40) were individually recruited from Amazon Mechanical Turk to participate as dyads (n = 20). Participants were paired with one another according to the order in which they began the experiment. All participants were over 18 years of age and fluent English speakers (self-reported), located within the U.S.

The experiment lasted an average of 13.19 minutes (range: 7.95—25.37 minutes). In return for their participation, all participants were paid $1.33 as base pay for finishing the experiment. Each participant also earned a bonus based of up to $2 for the entire experiment based on mean accuracy over all trials (mean = $1.80; range: $1.03—$1.95).

Procedure

All data collection procedures were completed through the experiment platform Dallinger (v3.4.1; `http://github.com/dallinger/Dallinger`), deployed on Amazon Mechanical Turk (`http://mturk.com`). Code for the experiment is available on GitHub (`http://github.com/thomasmorgan/joint-estimation-game`), and the resulting experiment data are available on the OSF repository for the project (`https://osf.io/8fu7x/`).

Each participant was individually recruited on Amazon Mechanical Turk to play a "Line Estimation Game" (advertisement: "Test your memory skills!"). Upon completing informed consent, participants were told that they would be playing a game in which they would be required to remember and re-create line lengths. Participants were informed that they would be complete their training trials individually and would then begin playing with a partner. Participants were given no information about their partner other than the guess that their partner made; no information about the partner's identity was shared.

In each trial, participants were shown 3 red lines (see figure; **NB**: add figure) and were asked to remember all three of them.[1] The 3 stimulus lines were displayed for 2 seconds then removed, providing participants with a blank screen for 0.5 seconds. Participants were then told which line to re-create (#1, #2, or #3) and were then given 1 second to submit their guess at how long the target line had been. To do so, participants were given a blank box and used their cursor to fill in the box with a blue line.

During training, participants were then shown the correct length of the target line (as a grey bar above their own guess) for 2 seconds. This was accompanied by a message telling the participant that they had guessed correctly ("Your guess was correct!") or incorrectly ("Your guess was incorrect") or that they had not submitted a guess within the 1-second time limit ("You didn't respond in time").

During testing, participants' stimulus viewing, waiting, and recreation times remained the same as during testing, but they no longer received information about whether their guess was correct or incorrect. Instead, after both participants had submitted their first guess, participants were shown their guess (in blue) above their partner's guess (in green). Both participants were then asked whether they wanted to change their own guess or to keep the guess they had submitted. If either participant in the dyad indicated that they wanted to change their guess, that participant was then allowed to change their guess (again with a 1-second time limit) *while* still being able to view their partner's guess. Participants who chose to keep their previously submitted guess was informed that their partner chose to submit a new guess and waited for the other participant to finish. At that point, participants were again allowed to change or keep their guess. This process continued until both participants chose to keep their guess.

Participants were informed that their final accuracy would only be calculated for their final guess. However, because they had no means to communicate with their partner about whether each would be accepting or changing their guesses, each participant could not have known whether their decision to keep the guess would have been their final guess for the trial.

[1]A pilot version of this study showed that participants adapted learned too quickly when given only 1 line to remember and recreate. The additional 2 lines were added to strictly increase the memory load, as opposed to adding difficulty in other ways (e.g., creating a moving stimulus).

For clarity, we will refer to each new stimulus set as a *trial* and to each submitted line length estimate within each trial as a *guess*. This means that some participants may have submitted multiple guesses per trial. The last submitted estimate—the one by which trial-level accuracy is calculated—will be referred to as the *final guess*.

All dyads completed 10 training trials (alone) and 15 test trials (with their partner). All training and test stimuli were randomly generated for each dyad, but both participants within the dyad were given the same stimuli. After participating, each individual participant was asked to complete a series of questionnaires about the game on a series of 1-10 Likert-style scales, including the perceived difficulty of the task, how engaged they were in the task, and questions about their own and their parner's cooperativeness and trustworthiness.

Analyses

Similarity To measure whether participants became more similar over time, we calculated the cross-correlation coefficient of participants' guess errors across trials (Paxton & Dale, 2013), within a window of +/-5 guesses. Although cross-correlation produces information about leading and following behavior, we have no *a priori* expectations about which of the two participants would emerge as a leader (given they have no information about their partner nor any assigned roles), so our first-pass analyses ignore any directionality by averaging across the each incremental lags (i.e., leading/following in both participants' directions).

Accuracy

Results

All analyses were performed in R (R Core Team, 2016). Linear mixed-effects models were performed using the `lme4` package (Bates, Mchler, Bolker, & Walker, 2015) using the maximal random slope structure for each random intercept to achieve model convergence (Barr, Levy, Scheepers, & Tily, 2013).

Discussion

Conclusion

Acknowledgements

Thanks go to the Dallinger development team for their assistance in executing the experiment.

This work was funded in part by the **DALLINGER GRANT INFO GOES HERE**.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure in mixed-effects models: Keep it maximal. *Journal of Memory and Language*, *63*(3), 255–278.

Bates, D., Mchler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of*

Statistical Software, *67*(1), 1–48. http://doi.org/10.18637/jss.v067.i01

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?, *6*, 3–5. http://doi.org/10.2307/41613414

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*, 59. http://doi.org/10.3389/fpsyg.2011.00059

Goldstone, R. L., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in Cognitive Science*, *8*, 548–568. http://doi.org/10.1111/tops.12212

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). PsiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842.

Paxton, A., & Dale, R. (2013). Argument disrupts interpersonal synchrony. *Quarterly Journal of Experimental Psychology*, *66*(11). http://doi.org/10.1080/17470218.2013.853089

Paxton, A., & Griffiths, T. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior Research Methods*, *49*(5), 1630–1638. http://doi.org/10.3758/s13428-017-0874-x

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/