## correspondence

# Disengagement with cognitive tasks decreases effect sizes

To the Editor — We appreciate the opportunity to comment on the Social Science Replication Project (SSRP) replication study<sup>1</sup>. The replication of our work<sup>2</sup> was the only project to yield a drastically smaller effect size in the first compared with the second sample (Cohen's d of 0.18 versus 0.69). When combined, the two samples independently replicated one of our major discoveries<sup>2</sup>, and in this sense bolstered the conclusions made in that work and subsequent research line<sup>3,4</sup>. But what should we conclude from the inconsistencies across samples? Were we merely the unlucky recipient of a false negative in the first sample? Or could another factor be at play? Fortunately, the SSRP team strived for transparency and author inclusion, so we were given access to the replication dataset. Our exploration revealed substantially lower task engagement in the replication compared with our original data, particularly in sample 1 (sample 1: 52%; sample 2: 57%; original: 66% correct; chance level ~33%). Subjects simply were not performing at the same levels as in our work, randomly guessing on a high proportion of trials.

Could measurement error explain the reduced and unstable effect size? At first blush, it seems intuitive that measurement error should hinder replications — how can you replicate a memory manipulation if participants do not engage in the memory task? At the extreme, this seems as fruitless as replicating a visual perception experiment with blindfolded participants. However, the replicating team dismissed our concerns, arguing that experimental error is just as likely to increase as decrease effect sizes. They were likely referring to the recent and somewhat counterintuitive work by Loken and Gelman<sup>5</sup>, showing that measurement error can inflate effect sizes under certain conditions.

These conditions, though, differ from the replication in two ways. First, the replication results were not selected based on statistical significance. Loken and Gelman show that measurement error increases effect size variability, particularly in small samples<sup>5</sup>. Increased variability can then inflate estimates when only reporting those that reach statistical significance. Conversely, without this sampling, measurement error generally decreases effect sizes, even in small samples.

The second difference is the type of measurement error introduced. In Loken and Gelman's simulations<sup>5</sup>, error was introduced by adding noise (that is, randomly increasing or decreasing observations) to two correlated variables. Like many cognitive psychology experiments, however, we measure task accuracy — an index that depends on making particular responses on particular trials. Adding measurement error in this context should systematically decrease accuracy, as seen in the replication of our work. To confirm this intuition, we simulated measurement error by increasing the proportion of trials on which participants made random responses (details available at https://osf.io/f24ug). Using

settings that closely mirror the original and replication datasets, increased measurement error reduced effect sizes by an alarming 20–35%. We also discovered more drastic effects in paradigms with fewer observations per subject.

In summary, we strongly recommend that data quality be taken into account when interpreting effect sizes and when planning replications. This may be particularly important for experimental protocols that assess complex indices (such as accuracy) because measurement error can have substantial and systematic consequences in these settings.

### Katherine Duncan<sup>1\*</sup> and Lila Davachi<sup>2</sup>

<sup>1</sup>Department of Psychology, University of Toronto, Toronto, Ontario, Canada. <sup>2</sup>Department of Psychology, Columbia University, New York, NY, USA.

\*e-mail: duncan@psych.utoronto.ca

#### Published online: 27 August 2018 https://doi.org/10.1038/s41562-018-0409-1

#### References

- Camerer, C. F. et al. Nat. Hum. Behav. https://doi.org/10.1038/ s41562-018-0399-z (2018).
- Duncan, K., Sadanand, A. & Davachi, L. Science 337, 485–487 (2012).
- 3. Patil, A. & Duncan, K. Psychol. Sci. 29, 45-55 (2018).
- Duncan, K. D. & Shohamy, D. J. Exp. Psychol. Gen. 145, 1420–1426 (2016).
- 5. Loken, E. & Gelman, A. Science 355, 584-585 (2017).

#### **Competing interests**

The authors declare no competing interests.