Sifting Signal From Noise With Replication Science

Chandra Sripada, Daniel Kessler, John Jonides

Science advances when signal is distinguished from noise, and an individual study is usually not capable of achieving reliable separation. We thus thank the Perspectives on Psychological Science Replication Initiative, the 23 contributing labs, and especially Martin Hagger for undertaking a Registered Replication Report (RRR) for the Sripada, Kessler, and Jonides (2014) protocol. This was a substantial endeavor and advances the field.

Our study was chosen for the replication initiative because we used a common task to induce depletion, the letter 'e' cancelling task, but adapted it for use with stimulus presentation software. This allows greater standardization and more detailed performance metrics. In our study, we observed a moderate-sized effect of depletion, Cohen's d = 0.69, which is comparable to many prior studies using the letter 'e' cancelling task. The Registered Replication Reports (RRR) results showed a much smaller effect. How should we interpret this finding?

We can start with a pessimistic interpretation. The RRR found that reaction time variability (RTV), the main dependent measure in our study, exhibited a Cohen's d of 0.04, which did not statistically differ from zero. We performed an additional Bayesian analysis based on the "Bayes Factor Test for Replication Success" (with associated R code) from Verhagen & Wagenmakers (2014). Specifically, we assessed whether the RRR result for RTV is more likely to be observed under: 1) the null hypothesis that there is no effect of our paradigm; or 2) a "replication" hypothesis. This hypothesis is formulated as a posterior distribution formed after observing the results from our original study. (Note: For all Bayesian analyses, the RRR per-subject results across multiple sites were collapsed together as if they arose from a single site as was done in Wagenmakers et al (2015). This effectively assumes that subjects are exchangeable across sites, and thus the Bayesian analyses should be interpreted with caution. Our R scripts are available from https://osf.io/47wfu/) This test yielded a Bayes factor of 0.013, i.e., the RRR result is 76.3 times more likely under null hypothesis than the replication hypothesis (See Figure 1, left panel), which constitutes "strong" evidence against the replication hypothesis. Based on this result, it would be reasonable to conclude that our paradigm produces no effect.

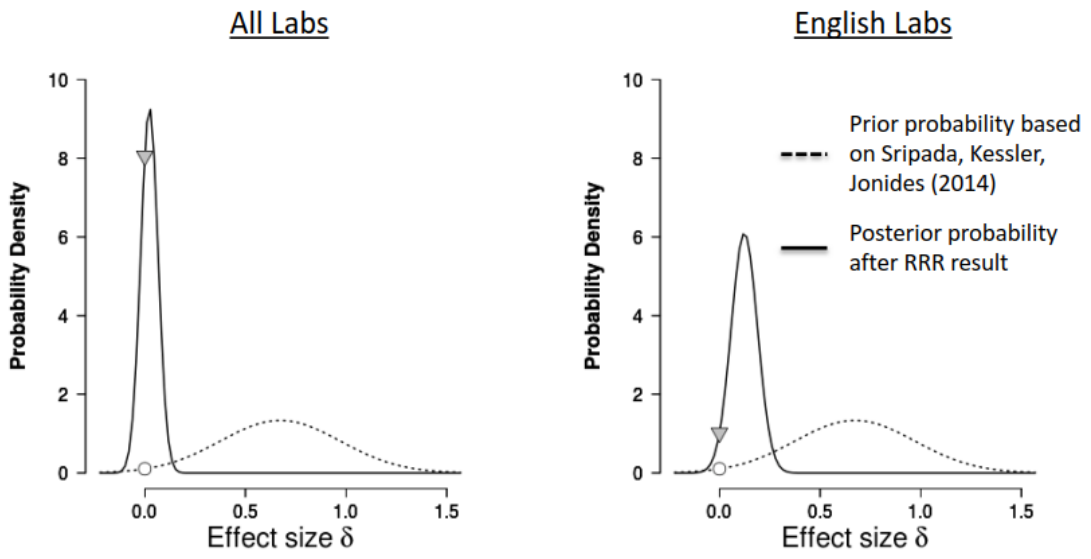

Figure 1*: Bayesian prior and posterior distributions for All (k=23) and English (k=11) labs. We performed a "Bayes Factor Test for Replication Success" based on Verhagen & Wagenmakers (2014), pooling across sites as in Wagenmakers et al (2015). The circles represent the prior probability of the null hypothesis while the triangles represent the posterior probability of the null hypothesis after observing the RRR result. The Bayes factors reported in the main text are represented in each panel as the height of the circle divided by the height of the triangle.*

A less pessimistic interpretation is also possible. The RRR involved 23 labs. Eleven labs (*n* after exclusions = 894) conducted the protocol in English and thus used our exact stimuli. Twelve labs conducted the protocol in other languages, and the RRR, with substantial assistance from one of us (DK), used standardized methods to try to make these stimuli as consistent as possible in terms of word length, word familiarity, etc. (stimuli are available here: https://goo.gl/rN3lZg). Because English language status was preregistered as a candidate moderator variable, we examined RTV results separately based on this variable. In the English labs, there was a small effect in the predicted direction (d = 0.14, 95% CI -0.02, 0.30). This effect was trend level statistically significant (p = 0.10) and it is contained within the 95% CI of the original study. In non-English labs, the effect was in the wrong direction (d = -0.04, 95% CI -0.18 to 0.10). The magnitude of the difference between the English and non-English labs is not statistically significant (p=0.12), but it is nonetheless suggestive. These observations raise the possibility that there is a non-zero RTV effect associated with our protocol, albeit a very small effect, with language, or some variable correlated with language, potentially moderating the effect.

We repeated the Bayesian analysis described above restricted to the 11 English labs. This yielded a Bayes factor of 0.105, i.e., the null hypothesis is 9.6 times more likely than the replication hypothesis, which constitutes "positive" evidence against the replication hypothesis (see Figure 1, right panel). So even if the RRR found a small

effect in English speaking labs, this effect is sufficiently small that it constitutes a failure to replicate our original study.

Regardless of whether the RRR found no effect or a small effect, what might explain why our study found an effect that is substantially larger? One possibility is chance. We may have observed an effect in our study that is an outlier given the size of the true effect. However, differences between our study protocol and the RRR protocol are also worth considering.

First, our study was conducted in an outpatient psychiatry clinic, not a university lab. Second, our participants were drawn from a community sample and were paid $20 for participation, while nearly all the RRR studies used undergraduate participant pools and did not pay participants. Third, our paradigm was implemented in the context of a double-blind drug manipulation with Ritalin and placebo. Participants arrived roughly 90 minutes before the tasks, were presented with an extensive consent detailing effects of Ritalin, ingested an unlabeled capsule, and then waited an hour for the drug (or placebo) effect to take hold. Given the preceding differences, it is likely that our participants were more motivated and invested in study participation than typical participants in the RRR studies. At least some models of the depletion effect (e.g., Inzlicht & Schmeichel, 2012) suggest sufficient motivation and investment during the first task is needed to observe self-control decrements in the second. It is also notable that we excluded roughly 13% of our participants due to inaccurate responding during the tasks, while the RRR studies excluded an average of 26% (33% in the English labs). This disparity in exclusions might be attributable to some of the differences identified above.

Finally, the tasks that have thus far been used in the ego depletion literature are heterogeneous and there is a need to better understand to what extent the specific component processes involved exhibit "depletability." Our two computerized tasks involve a number of component processes: response inhibition (inhibiting a prepotent motoric response), interference processing (registering a stimulus dimension while avoiding distractors), and sustained attention (maintaining task-directed focus for extended periods of time), among others. Interestingly, sustained attention has reliably been found to exhibit the vigilance decrement effect—a depletion-like effect—in an entirely independent literature that spans over fifty years of research (Mackworth, 1948; Davies & Parasuraman, 1982; See, Howe, Warm, & Dember, 1995). Vigilance decrement is usually exhibited over longer timescales (> 30 minutes), suggesting that investigating tasks of longer duration could be a fruitful approach going forward (see also Lee, Chatzisarantis, & Hagger, 2016). More broadly, it is likely the component processes engaged by our tasks differ in important respects from those engaged in the highly varied tasks in other studies of ego depletion. Caution is thus required in drawing implications from the results of this RRR for the phenomenon of ego depletion writ large.

References

Davies, D. R., & Parasuraman, R. (1982). *The Psychology of Vigilance*. New York: Academic Press.

Inzlicht, M., & Schmeichel, B. J. (2012). What Is Ego Depletion? Toward a Mechanistic Revision of the Resource Model of Self-Control. *Perspectives on Psychological Science*, *7*, 450–463.

Lee, N., Chatzisarantis, N., & Hagger, M. S. (2016). Adequacy of the Sequential-Task Paradigm in Evoking Ego-Depletion and How to Improve Detection of Ego-Depleting Phenomena. *Frontiers in Psychology*, 136.

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *The Quarterly Journal of Experimental Psychology*. doi:10.1080/17470214808416738

See, J. E., Howe, S. R., Warm, J. S., & Dember, W. N. (1995). Meta-analysis of the sensitivity decrement in vigilance. *Psychological Bulletin*, *117*, 230–249.

Sripada, C., Kessler, D., & Jonides, J. (2014). Methylphenidate Blocks Effort-Induced Depletion of Regulatory Control in Healthy Volunteers. *Psychological Science*, 0956797614526415.

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475.

Wagenmakers, E.-J., Verhagen, J., & Ly, A. (2015). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, 1–14.