No relationship between researcher successful productivity and replicability: an analysis of four studies with 79 replications.

John Protzko, Jonathan Schooler

University of California, Santa Barbara

Abstract

What explanation is there when teams of researchers are unable to successfully replicate already established ‘canonical’ findings? One suggestion put forward, but left largely untested, has been that those researchers who fail to replicate other studies are of lower ‘caliber’, lacking the expertise and skill necessary to successfully replicate such experiments. Here we empirically test the validity of those claims across 79 laboratories of differing ‘caliber’ replicating four different studies. Using a bibliometric tool as our indicator of ‘caliber’, we find absolutely no empirical evidence for the researcher ‘caliber’ and reproducibility hypothesis. Claims are now being put forward to explain replication failures in psychological science. These hypotheses carry a lot of potential for explaining scientists’ behavior. The results, however, do not uphold the hypotheses; alternate explanations instead should be sought for explaining replication failure.

Keywords Reproducibility; Bibilometrics; Metascience, h-index, Registered Replication Reports

Introduction

Science is not a solipsistic pursuit. The purpose is to discover truths about the world using methods that other scientists can reproduce. The validity of a discovery is open to testing by other researchers. When only the 'discovering' lab can reproduce a result, trust in the finding can and should be examined.

With a new vigor towards reproducibility, psychologists have now started to direct effort at testing the replicability of given research findings. The results of these large-scale replication attempts have introduced new questions into the field. One such initiative ran single replications of 100 studies and found less than half of the studies replicated (OSC, 2015). While the exact rate has been debated (e.g. Gilbert et al., 2016a-b), the question of *why* such a difference in reproducibility rates has been asked (Anderson et al., 2016).

There are many reasons why a given research finding may not replicate. Among the more nefarious include the original data was fraudulent and thus, the finding was not real (see John et al., 2012 for researcher rates of admitting to this). A little less nefarious is researchers engaging in selective reporting, dropping of conditions or participants, or freedom in analysis to make findings 'appear' (e.g. Simmons et al., 2011). Both of these reasons for irreproducibility arise from the original findings being nonexistent. Thus, there is nothing to reproduce.

Given that the original findings are true, however, what could account for differences in reproducibility? Statistical reasons include power (Cohen, 1969) and fidelity to experimental procedures (see Gilbert et al., 2016b). In this manuscript, we investigate another possible explanation, researcher 'caliber'.

Research, especially involving studies with deception, can be difficult to conduct. To remove as much systematic error as possible, interactions with participants are heavily scripted. To successfully pull off deception or accurate measurement/assessment in the name of research, expertise may be required. It has been suggested that researchers who engage and fail in replicating canonical studies are of inferior caliber (Bartlett, 2014; Cunningham & Baumeister, 2016). As the hypothesis has been put forward that researchers of different caliber are better or less able to experimentally replicate an effect, we sought to answer the question empirically.

Replications

To test the hypothesis that reproducibility is a function of researcher 'caliber', we collected 79 replications that had been conducted of four studies. We used the four published Registered Replication Reports (RRRs), investigations where a dozen or more individual research teams all attempt to replicate the same study.

Registered Replication Reports are direct replications of a study conducted by multiple, independent laboratories of researchers who vary in the extent to which they believe in the original finding. All labs follow the exact same protocol that is approved by the original study team before data collection begins.

The reason for using this as our sample was that it provided multiple replications of the same basic effect by researchers of varying levels of expertise. In one replication investigation (OSC, 2015), researchers who had more publications chose more robust studies to offer to replicate (Bench, et al., 2017). After taking this initial volunteering into effect, there was no residual relationship between researcher 'caliber' and replication success (cf. Cunningham & Baumeister, 2016). As that investigation only looked at one replication per study, however, it was unable to look at variation within the same study. The analysis proposed here is able to look

at variation in ‘caliber’ within multiple replications of the same study. Thus, instead of one effect having one replication and averaging across different studies, each study under replication has multiple effect sizes. This provides the strongest test of the ‘researcher caliber and reproducibility’ hypothesis.

The four RRRs represent multi-lab replications of the following phenomena: verbal overshadowing (Schooler & Engster-Schooler, 1991); priming commitment and reaction to hypothetical romantic betrayal (Finkel et al., 2002); the facial feedback hypothesis (Strack et al., 1988); and ego depletion (Hagger et al., 2015).

Materials and Methods

Verbal Overshadowing RRR

The original finding under investigation was that verbally describing the face of someone causes a decrease in their ability to accurately point out the face in a lineup (Schooler & Engster-Schooler, 1991). 23 separate labs engaged in a replication of this study (Alonga et al., 2014; study 2). The 23 labs were able to successfully replicate the result. While the test statistic for heterogeneity was not significant (Q (21) = 15.25, p > .809), a separate investigation by us revealed that the 95% confidence interval for degree of heterogeneity was 0% to 46%. Therefore, there exists the possibility that differences in researcher ‘caliber’ can explain the variation in effect sizes.

Priming Commitment RRR

The original finding under investigation was that inducing commitment in relationships causes people to be more forgiving of hypothetical betrayals (Finkel et al., 2001, study 1). 16 separate labs engaged in a replication of this study (Cheung et al., 2016). There were a possible

four dependent variables used. We chose the one most consistent with the hypothesis and showed the largest amount of potential heterogeneity (neglect responses) for investigation here. These neglect responses are passive ways of trying to undermine a relationship, such as giving someone the 'cold shoulder.' Using this dependent variable maximizes the chance of testing our hypothesis. The 16 labs were unable to successfully replicate the commitment priming on neglect responses. While the test statistic for heterogeneity was not significant (Q (15) = 18.09, p > .257), a separate investigation by us revealed that the 95% confidence interval for degree of heterogeneity was 0% to 54%. Therefore, there exists the possibility that differences in researcher 'caliber' can explain the variation in effect sizes.

Facial Feedback RRR

The original finding under investigation was the forerunner of embodied cognition, making a smile by holding a pencil upwards in your mouth altered the way the brain interprets humorous videos (Strack et al., 1988, study 1). 17 separate labs engaged in a replication of this study (Wagenmakers et al., 2016). The 17 labs were unable to successfully replicate the result. The RRR claimed there was no heterogeneity within their meta-analysis. A separate investigation by us revealed that the 95% confidence interval for degree of heterogeneity was 8% to 70%. Therefore, there exists the possibility that differences in researcher 'caliber' can explain the variation in effect sizes.

Ego Depletion RRR

The original finding under investigation was that engaging in a difficult cognitive control task depletes people's resources and hampers performance on a subsequent task (reaction time variability; Sripada et al., 2014). 23 separate labs engaged in a replication of this study (Hagger

et al., 2016). The 23 labs were unable to successfully replicate the result. While the test statistic for heterogeneity was not significant (Q (22) = 20.12, p > .575), a separate investigation by us revealed that the 95% confidence interval for degree of heterogeneity was 0% to 45%. Therefore, there exists the possibility that differences in researcher 'caliber' can explain the variation in effect sizes.

Researcher 'caliber'

We used the h-index—a metric of researcher caliber/impact/experience (Hirsch, 2005)—of the researchers who undertook replications of the various effects. The h-index for a given scientist is a function of the number of papers that author has published and the number of times those papers have been cited. Thus, it is more sensitive to research impact than simply measuring the raw number of publications. Typical h-indices for psychological scientists range from six to ten (Ruscio et al., 2012). Previous research into the replicability success of different research teams have used number of publications as a measure of "high-expertise" (Cunningham & Baumeister, 2016, p. 12; citing Bench et al., 2017). The h-index also incorporates a function of the number of times that research has been cited. Thus, it incorporates impact within a field and the number of impactful papers a researcher has published. It is a better metric of researcher 'caliber' than simply number of publications, which may more likely be a metric of solely researcher productivity (Hirsch, 2005).

Overall analyses

To test the researcher 'caliber' and reproducibility hypothesis, we collected the raw effect sizes from each of the four above-mentioned RRR. We then matched the obtained effect size with the lead authors of the replication's h-index. These h-indexes were collected from both Web

of Science and GoogleScholar between October 3-17, 2016. To test the hypothesis, we used meta-regression with the obtained effect size, weighted by the meta-analytic standard error for each study, testing whether researcher 'caliber' was associated with the obtained effect size. There was sufficient variation in h-indexes (0 to 54) to allow for testing this prediction. All analyses were run in STATA 13.1. We first present the analyses for each individual RRR, followed by a pooled analysis across all four.

Results

Verbal Overshadowing

There was no evidence for the researcher 'caliber' and reproducibility hypothesis. Specifically, more experienced researchers were just as likely to return the same effect size as novice researchers (b = -.003, p > .671). Thus, while verbal overshadowing was successfully replicated, variation in the magnitude of the effect was not related to researcher 'caliber' (see Fig. 1).

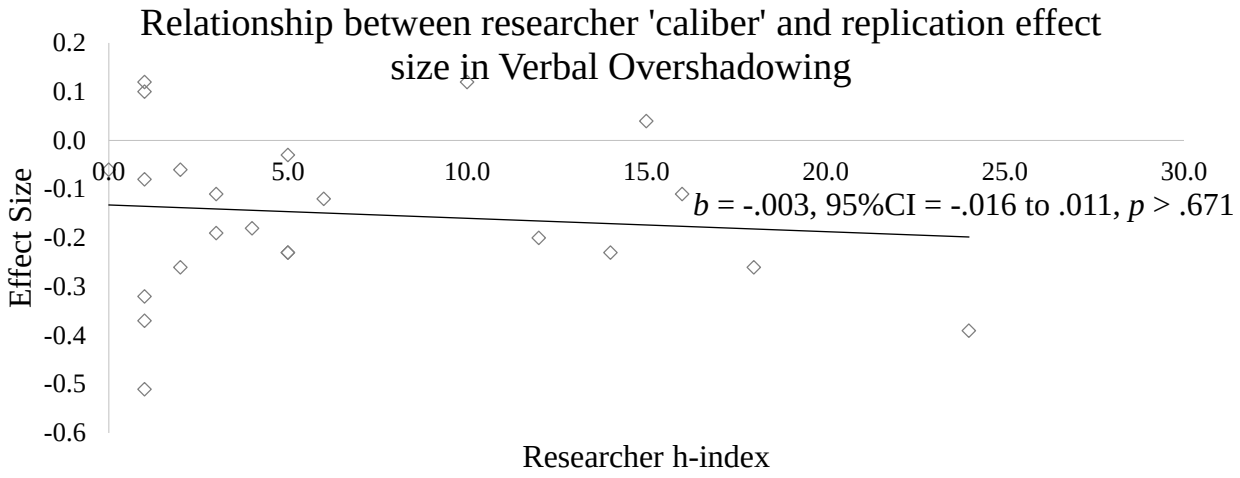

Figure 1: Meta-regression of researcher 'caliber' on obtained effect size in replicating the verbal overshadowing paradigm.

Priming Commitment

There was no evidence for the researcher ‘caliber’ and reproducibility hypothesis. Specifically, more experienced researchers were just as likely to return the same effect size as novice researchers (b = -.008, p > .654). Within the different laboratories attempting the replication, those who returned an effect consistent with the original hypothesis were of no different ‘caliber’ than those who did not (see Fig. 2).

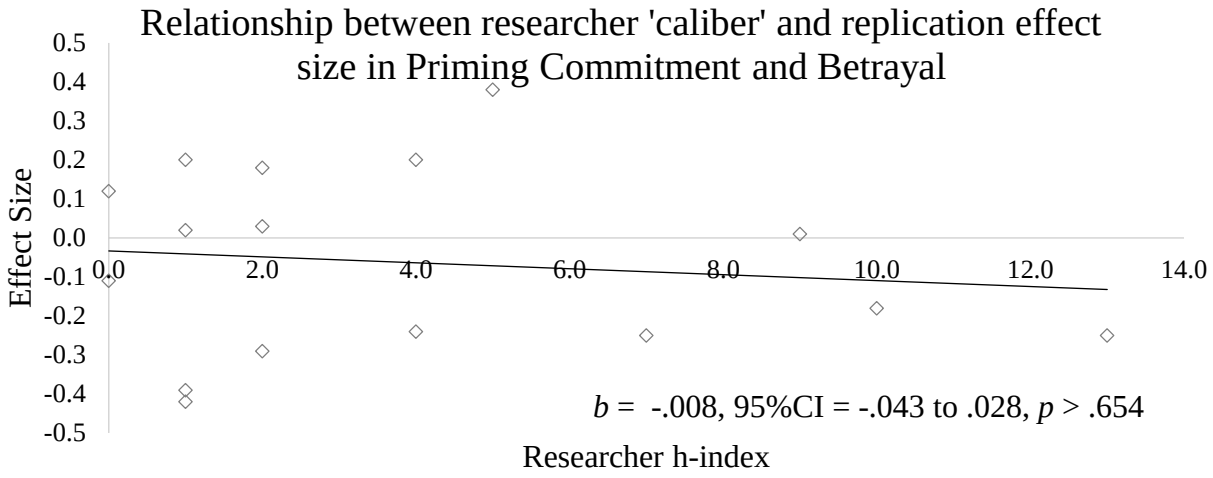

Figure 2: Meta-regression of researcher 'caliber' on obtained effect size in replicating the priming commitment and betrayal paradigm.

Facial Feedback

There was no evidence for the researcher ‘caliber’ and reproducibility hypothesis. More experienced researchers were just as likely to return the same effect size as novice researchers (b = .005, p > .283). The overall result of the RRR was unable to replicate the Facial Feedback hypotheses. Within the different laboratories attempting the replication, those who returned an effect consistent with the original hypothesis were of no different ‘caliber’ than those who did not (see Fig. 3).

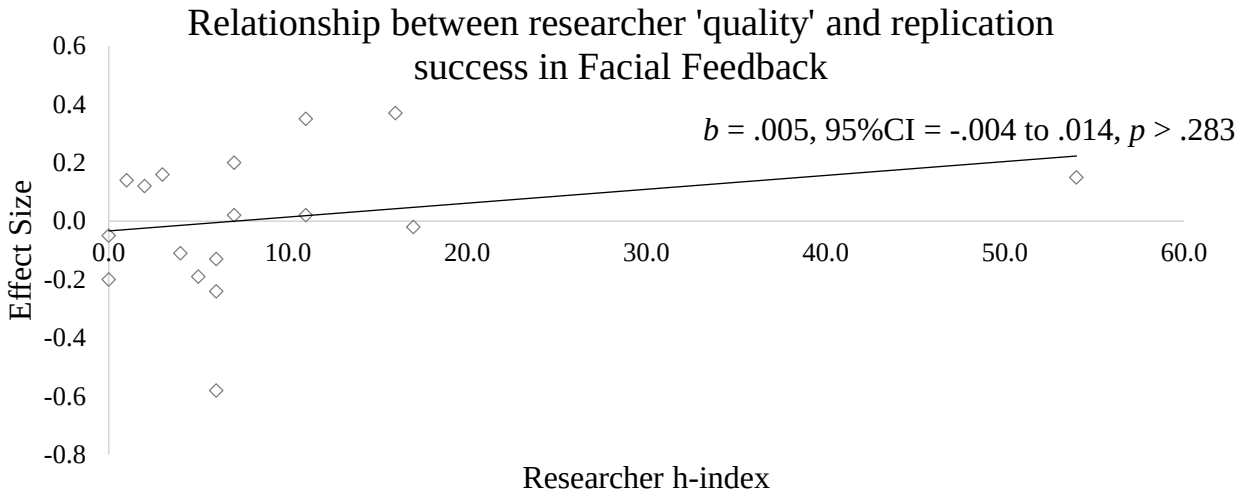

Figure 3: Meta-regression of researcher 'caliber' on obtained effect size in replicating the facial feedback paradigm.

Ego Depletion

The ego depletion RRR presented a break from the other three pattern of results. Unlike before, we did observe a relationship between researcher ‘caliber’ and the observed effect of exerting a large amount of mental control decreasing performance on a subsequent task. The results, however, ran *counter* to the researcher ‘caliber’ and reproducibility hypothesis. Better ‘caliber’ researchers actually observed smaller, indistinguishable from zero effects (failing to replicate), whereas more novice researchers observed larger effects of depletion (b = -.106, p < .044; see Figure 4).

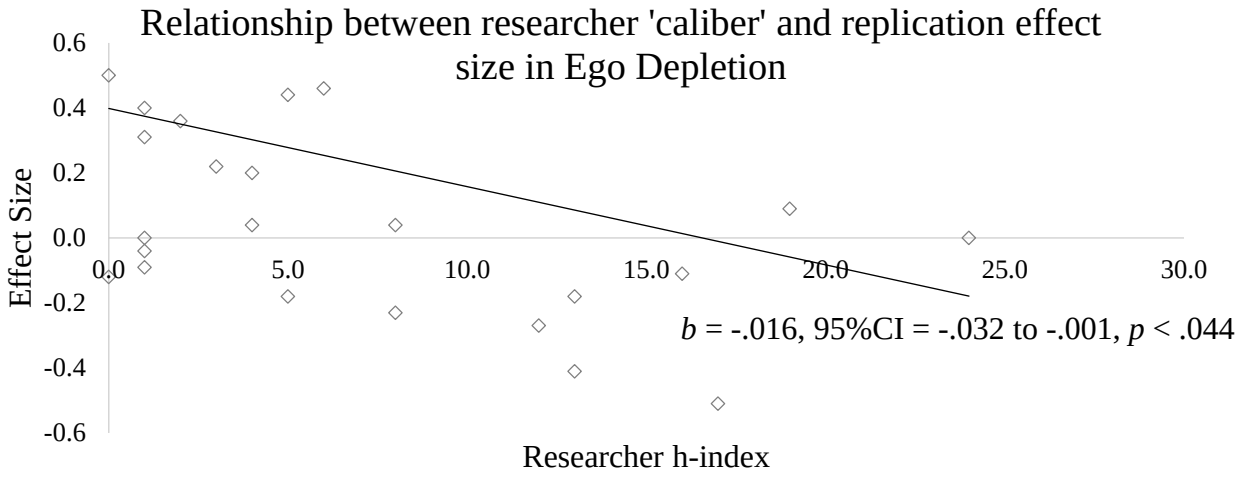

Figure 4: Meta-regression of researcher 'caliber' on obtained effect size in replicating the ego depletion paradigm.

Though against the hypothesis, this finding puts one into a bind. Either: a) ego depletion by this paradigm is not a real effect and we have evidence that higher 'caliber' researchers can confirm that; or b) ego depletion by this paradigm is a real finding but it can only be found by novice—low 'caliber'—researchers (cf. Cunningham & Baumeister, 2016). There is also the possibility that this is a statistical fluke, which may be the most likely given the consistent null results of the other RRRs. We take no firm stance here and believe future replication efforts of ego depletion using multiple labs of a variety of expertise should be conducted using a different paradigm to elucidate this result.

Overall Results

In each of the four RRRs reported so far there was no evidence for the researcher 'caliber' and reproducibility hypothesis. It is possible that within each paradigm the effect was

underpowered. Therefore, as a robustness check, we combined all four RRRs and tested once again the hypothesis that returned effect size in a replication is a function of the experience the researcher brings to the experiment.

In RRRs where the original result was negative in value, confirmation of the researcher 'caliber' hypothesis would indicate a negative slope to the regression of 'caliber' on effect size, with 'better' researchers contributing effect sizes further away negatively from zero (Verbal Overshadowing; Priming Commitment). In RRRs where the original result was positive, however, the predicted slope would be positive, with larger effect sizes from higher 'caliber' researchers (Facial Feedback; Ego Depletion). As this combination of positive and negative slopes could cancel each other out, we switched the sign of the effect sizes from Verbal Overshadowing and Priming Commitment so the overall prediction would indicate a positive slope.

The results confirmed the overall fact that there is no empirical support for the researcher 'caliber' and reproducibility hypothesis. There was no relationship between the 'caliber' of the researchers in these four RRRs and the absolute value of the effect sizes they returned (b = .0003, p > .992; see Fig. 5). One may believe that the researcher with an h-index of 54 represents a statistical outlier that may bias the results. It does not. Running the same meta-regression with that outlier removed does not change the results (b = -.002, 95%CI = -.009 to .006, p > .716). Experimenters who had not published a single study before were just as likely to converge on the same meta-analytic effect as those who had great skill running studies, and publishing papers that in turn are heavily cited.

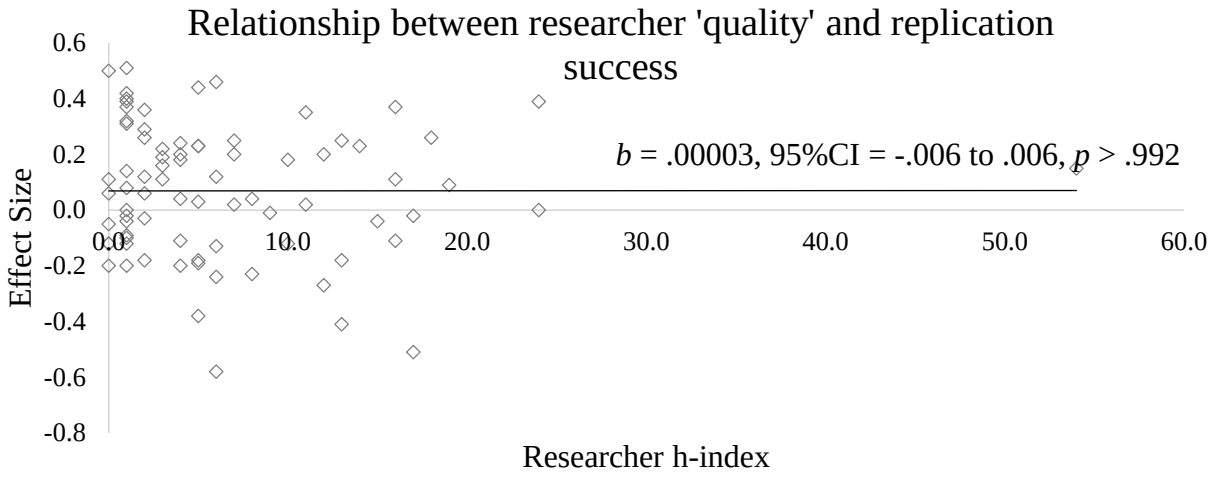

Figure 5: Meta-regression of researcher 'caliber' on obtained effect size across replications in all four RRRs.

Furthermore, to fully argue for a ‘no effect’ relationship, we also calculated the 90% confidence interval around the estimate (90%CI = -.005 to .005; see Lakens, *In Press*). Unless the researcher ‘caliber’ and reproducibility hypothesis would explicitly state that each unit increase in researcher ‘caliber’ should lead to a .004 increase in effect size (as that effect is not rejected by the data here) *and no higher*, we can conclude the results here allow for such a ‘no effect’ interpretation.

Effect sizes in these instances are a function of the mean difference between two groups divided by the precision of their estimate. It could be the case that researchers of higher ‘caliber’ get more precise estimates—not necessarily larger effect sizes. For this to be the case, however, there would need to be a *negative* relationship between ‘caliber’ and mean differences. Holding mean differences constant, if precision increased then effect sizes would go up. The only way for the ‘precision’ effect to be possible considering the above (largely) null effects on effect sizes would be smaller mean differences mixed with estimates that are more precise across ‘caliber’. A

supplementary meta-regression showed that there is no relationship between group mean differences and 'caliber' in any study (all *p*s > .385). Thus, there could not be a 'precision' effect—however measured—without showing either an effect size effect or a mean differences effect.

Discussion

The question of whether expertise in psychology predicts replication success is one that is important to understand when considering the implications of large-scale replication efforts. If more novice or lower 'caliber' researchers are unable to find basic canonical effects when replicating, it is extremely important to take such information into account.

Here we used a metric of researcher 'caliber' to look at replication success among labs conducting the same studies. This way, we had increased power to detect variation between labs. Our results showed no evidence whatsoever in favor of the researcher 'caliber' and reproducibility hypothesis. In three of the four RRRs, there was no association between obtained effect size and the 'caliber' of the researcher conducting the replication. In one of the RRRs, we actually saw evidence that more experienced researchers were closer to returning the overall meta-analytic effect of zero, with less experienced researchers being the ones who found evidence for ego depletion. Collapsing across all four RRRs, the relationship was zero (b = .00003, p > .992).

Thus, it did not matter whether the RRR successfully replicated the initial results. The very claim of the researcher 'caliber' and reproducibility hypothesis is that despite meta-analytic evidence to the contrary, the effect still exists. The fact that the aggregate of studies could not

find it is immaterial to the argument; high ‘caliber’ researchers should have been able to. The problem is, that was not the case. When the effect was faithfully replicated (Verbal Overshadowing), researcher ‘caliber’ did not matter. The only time it *did* alter results was in the Ego Depletion replication—where high ‘caliber’ researchers confirmed the null effect and novice researchers ‘found’ evidence for ego depletion.

We agree wholeheartedly with the sentiment that “metascience is not exempt from the rules of science” (Gilbert et al., 2016a, p. 1037-a). As such, we sought to empirically test the belief among some researchers that reproducibility is a function of the caliber/expertise of the one conducting the study, opposed to letting the hypothesis stand untested. Such a hypothesis deserves empirical attention as replication becomes more accepted within psychological science.

Since this investigation used multiple labs replicating the same studies, our results are uncontaminated by publication bias or missing studies. All research teams followed the exact same protocol with no deviations, so lack of fidelity to experimental task was minimal. All original authors approved the replication materials before the study began. There was sufficient variation among indices of researcher ‘caliber’, as well as sufficient variation of effect sizes to permit accurate testing. Furthermore, there was enough heterogeneity within each RRR to allow the possibility that differences in returned effect size could be a function of something other than random chance. Researcher ‘caliber’ was not it.

The RRRs we used ranged from those that successfully replicated the original finding (Alonga et al., 2014), to those that successfully replicated the manipulation but failed to provide evidence for the outcomes (Hagger et al., 2016), to those which were unable to replicate the basic manipulation (Cheung et al., 2016). Therefore, there was a range of possibilities that could have arisen, including different relationships between researcher ‘caliber’ and effect sizes in

different paradigms. That the only relationship to emerge was one backwards to the hypothesis under test (that higher 'caliber' researchers were *less* likely to replicate the result of the ego depletion study) stands as evidence against the researcher 'caliber' and reproducibility hypothesis.

Limitations

The h–index is not without its problems as a metric of researcher 'caliber' (e.g. Yong, 2014). There are few if any other objective measures that could be considered better, however. This could plausibly limit the generalization unless a better metric if found. The problem with defining researcher 'caliber' is avoiding circularity. If only high caliber researchers can replicate ego depletion, for example (see Cunningham & Baumeister, 2016), then how can we define 'high caliber' outside of 'is able to replicate ego depletion'? We believe the h-index provides an objective metric that combines number of publications and impact of those publications. As such, it avoids such circularity. Better metrics could be used in the future with further large-scale replication attempts to tease apart the nature of any relationship.

Furthermore, the RRRs are tightly controlled experiments where the editorial staff gives a lot of support to the experimenters; this may reduce the opportunity for caliber differences to arise. We believe this situation is unlikely, as our supplementary analysis of the mean differences and thus, 'precision' of the estimates also showed no effect. Furthermore, if the work of the editorial staff for the RRRs turns the novice replicators effectively on par with experts, then there is an immediate tension. The results from the individual RRRs cannot now be invalid because novices undertook them. Either the RRR meta-analytic results hold, or the effect represents what is true in the world (often null); or the effect exists as in the original study, but higher 'caliber' researchers cannot find it any better than lower 'caliber' ones.

Conclusion

The researcher 'caliber' and reproducibility hypothesis is an attractive one in interpreting large-scale failures to replicate. While it may be tempting to dismiss such a hypothesis outright, metascience is not exempt from empirical testing. Thus, we have directed the strongest empirical test of the hypothesis to date and found it lacking evidence.

Acknowledgements: We would like to thank Drs. Stephen Lindsay and Dan Simons for their comments on the argument laid out here and Dr. Daniel Lakens for suggesting the CI equivalence approach.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Bornstein, B., ... & Carlson, C. (2014). Contribution to Alonga et al (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science,* 9(5), 556-578.

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., ... & Della Penna, N. (2016). Response to comment on "estimating the reproducibility of psychological science". *Science*, 351(6277), 1037-1037.

Bartlett, T. (2014). Replication crisis in psychology research turns ugly and odd. *The Chronicle of Higher Education*, June, 23.

Cheung, I., Campbell, L., LeBel, E., . . . Yong, J. C. (2016). Registered Replication Report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, 11, 750–764.

Cohen, J. (1969). *Statistical power analysis in the behavioral sciences*. New York: Academic Press

Cunningham, M. R., & Baumeister, R. F. (2016). How to Make Nothing Out of Something: Analyses of the Impact of Study Sampling and Statistical Interpretation in Misleading Meta-Analytic Conclusions. *Frontiers in Psychology*, 7.

Finkel, E. J., Rusbult, C. E., Kumashiro, M., & Hannon, P. A. (2002). Dealing with betrayal in close relationships: Does commitment promote forgiveness? *Journal of Personality and Social Psychology*, 82, 956–974.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016a). Comment on "Estimating the reproducibility of psychological science. *Science,* 351(6277).

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016b). A RESPONSE TO THE REPLY TO OUR TECHNICAL COMMENT ON "ESTIMATING THE REPRODUCIBILITY OF PSYCHOLOGICAL SCIENCE".

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11, 546–573.

Higgins, J. P. T. & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21, 1539-1558.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102, 46, 16569-16572.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science,*

Lakens, D. (in press). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

Ruscio, J., Seaman, F., D'Oriano, C., Stremlo, E., & Mahalchik, K. (2012). Measuring scholarly impact using modern citation-based indices. *Measurement: Interdisciplinary Research and Perspectives*, 10(3), 123-146.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22, 36–71.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*,

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54, 768-777.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr.,…Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11