Journal of Experimental Social Psychology 68 (2017) 181–184

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

FlashReport

Does expertise matter in replication? An examination of the reproducibility project: Psychology☆

Shane W. Bench, Grace N. Rivera, Rebecca J. Schlegel *, Joshua A. Hicks *, Heather C. Lench *

Department of Psychology, Texas A&M University, 4235 TAMU, College Station, TX 77843-4235, United States

HIGHLIGHTS

- The issue of replication has received a large amount of recent interest.
- High expertise teams obtained effect sizes larger than low expertise teams.
- High expertise teams also selected original studies with larger effect sizes.
- The overall pattern of results suggested expertise mostly impacts study selection.
- Experts may consider different methodological criteria at study selection.

ARTICLE INFO

Article history:
Received 3 May 2016
Revised 12 July 2016
Accepted 12 July 2016
Available online 14 July 2016

Keywords:
Replication
Expertise
Experimentation
Social psychology
Cognitive psychology

ABSTRACT

A recent article reported difficulty in replicating psychological findings and that training and other moderators were relatively unimportant in predicting replication effect sizes. Using an objective measure of research expertise (number of publications), we found that expertise predicted larger replication effect sizes. The effect sizes selected and obtained by high-expertise replication teams was nearly twice as large as that obtained by low-expertise teams, particularly in replications of social psychology effects. Surprisingly, this effect seemed to be explained by experts choosing studies to replicate that had larger original effect sizes. There was little evidence that expertise predicted avoiding red flags (i.e. the troubling trio) or studies that varied in execution difficulty. However, experts did choose studies that were less context sensitive. Our results suggest that experts achieve greater replication success, in part, because they choose more robust and generalizable studies to replicate.

The Open Science Collaboration (2015) recently published a report on the reproducibility of psychological science. The results were troubling, with only 36% of the replications producing a statistically significant result and replication effect sizes that were half the magnitude of the original effect sizes. This formidable project involved the attempted replication of 100 studies in social and cognitive psychology, and brought home the impact and scope of numerous problems in methodological practice that could inflate effect sizes. However, the studies were conducted by multiple teams of researchers who had highly variable experience in conducting experiments. Psychologists study for years under experienced mentors, learning the skills necessary to successfully conduct studies. Although the importance of expertise has been noted for decades (Smith, 2000; Wilson, Aronson, & Carlsmith, 2010), and recently discussed in both news and social media (Carey, 2015; Schimmack, 2015), there has been no empirical evidence that expertise matters. Just as master chess players and seasoned firefighters develop intuitive expertise that aids their decision making (Kahneman & Klein, 2009), seasoned experimenters may develop intuitive expertise that influences the "microdecisions" they make about study selection (e.g., what study to conduct, what measures to use), and data collection (e.g., who interacts with participants, where data collection occurs, how to manage unexpected problems). These "microdecisions" could affect experimental control and thus replication effect size.

The OSC data provides a rare opportunity to explore these ideas. The original publication (OSC, 2015) included self-reported expertise and self-reported number of publications and reported that there was no relationship between these indicators and various indicators of replication success. A check of this data, however, indicated potential discrepancies in authors' reports and objective indicators of publications (e.g., some reports did not match database information). Accordingly, we independently collected this data and explored whether replication team expertise predicted replication effect size. We a priori hypothesized that more experienced replication teams would obtain larger replication effect

☆ Thanks to the Templeton Foundation working group on Virtue, Happiness, and Meaning in Life for comments and discussion to the last author.

* Corresponding authors.
E-mail address: hlench@tamu.edu (H.C. Lench).

http://dx.doi.org/10.1016/j.jesp.2016.07.003
0022-1031/

sizes than less experienced teams. Further, based on data patterns, we conducted exploratory analyses to examine whether experts also select studies with larger effect sizes to replicate. That is we tested whether expertise mattered *both* in the selection of studies to replicate and in the execution of replication studies.

1. Method

The original 100 studies included in the OSC (2015) report, minus the three studies they excluded, were examined. We conducted a PsycInfo search for the first author and the senior author identified in the replication reports. We recorded total number of publications, number of publications containing data, and number of publications including experiments (the latter two characteristics identified from abstracts; all indicators were highly correlated, $r_{number*data} = 0.99$, $r_{number*experiment} = 0.61$, $r_{data*experiment} = 0.56$, p's < 0.001).

For analyses, we used the *r* effect sizes from the OSC report. The program Comprehensive Meta-Analysis (CMA; Version 3; Biostat, 2014) was used to compare effect sizes. We changed the direction of two effects that were coded incorrectly in the OSC data upon checking the replication reports (e.g., a study coded as a negative effect when the finding was consistent with hypothesized direction). We evaluated the relationships between objective expertise and effect sizes using the meta-regression program within CMA with a maximum likelihood procedure and a fixed effects model (Bornstein & Cooper, 2009; Higgins & Thompson, 2002). The methods and hypotheses for analyses related to replication effect sizes were preregistered on the Open Science Framework (osf.io/mn2zd; analyses focused on original effect sizes were not preregistered because they were not anticipated).

2. Results

2.1. Primary analyses: expertise effects on study replication

Our primary hypothesis was that the experience of the replication team would predict replication effect size. We conducted meta-regression analyses with effect size predicted by the three indicators of combined expertise of the first and senior author. The number of publications by the replication team significantly predicted replication effect sizes, with greater effect sizes obtained by teams that had published more, $b = 0.0011$ ($SE = 0.0002$), 95% CI [0.0007, 0.0015], $Q_R = 24.42$, $p < 0.0001$. Although the coefficients were small, the regression lines captured the data points well (as evidenced by the low standard errors) and are statistically significant, suggesting that greater expertise predicted greater replication effect sizes. Number of publications with data and number of publications with experiments also predicted replication effect size (p's < 0.001; see Table 1 on osf.io/7wgk9 for these and supplemental analyses).

The continuous analyses suggested that the expertise of the research team mattered. However, there was a broad range of expertise present (range from 0 to 203 publications, $SD = 39.48$) and many replication teams had no publications (11%). This resulted in an unanticipated and highly skewed distribution (skew = 3.21), with a large cluster of scores at or near zero (kurtosis = 11.09; see Fig. 1). This distribution violated assumptions of inferential statistics that depend on mean, raw, or rank order scores.

To provide a more robust estimate of the impact of expertise, expertise was split into those replication teams with fewer than ten publications (52 teams) and those with ten or more publications (a value that is near the median and approximates the expected starting level for a tenure-track position, 45 teams).

As shown in Fig. 2, a mixed effects analysis revealed that the effect sizes obtained by high-expertise teams was nearly twice as large ($r = 0.25$, 95% CI [0.161, 0.333], $Z = 5.41$, $p < 0.001$) as those obtained by low-expertise teams ($r = 0.14$, 95% CI [0.101, 0.161], $Z = 6.52$, $p < 0.001$), $Q_B(1) = 4.42$, $p = 0.036$ (analyses with other indicators of expertise are reported in Supplemental materials). This difference remained after controlling for self-reported domain and methodological expertise (see Table 3 in Supplemental materials). The impact of expertise (Fig. 2) was more pronounced for social psychology studies ($r_{highexpertise} = 0.18$, 95% CI [0.068, 0.296]; $r_{lowexpertise} = 0.09$, 95% CI [0.046, 0.138]; $Q_B(1) = 2.12$, $p = 0.146$) relative to cognitive psychology studies ($r_{highexpertise} = 0.34$, 95% CI [0.207, 0.462], $r_{lowexpertise} = 0.28$, 95% CI [0.176, 0.379]; $Q_B(1) = 0.52$, $p = 0.473$).

Taken together, these results suggest that high expert teams obtain larger effect sizes than less expert teams. This is consistent with our original hypothesis and the idea that expertise may influence the "microdecisions" scientists make as they conduct replication studies. However, supplementary exploratory analyses suggest a more nuanced conclusion.

2.2. Exploratory analyses: expertise effects on study selection

2.2.1. Original effect sizes

We conducted meta-regression analyses with original study effect size predicted by the three indicators of expertise used in the primary analyses. The number of publications by the replication team significantly predicted original effect sizes, indicating that teams that had published more selected studies to replicate that had greater original effect sizes, $b = 0.0034$ ($SE = 0.0002$), 95% CI [0.0030, 0.0038], $Q_R = 245.86$, $p < 0.0001$. Comparing the Q_R between replication and original effect sizes, which approximates the magnitude of the relationship between expertise and effect sizes, reveals a markedly stronger impact of expertise on original effect sizes than replication effect sizes. Number of publications with data and number of publications with experiments also

Table 1
Characteristics of studies selected by high versus low expertise replication teams.

	t-Value	*p*-Value	Cohen's *d*	High experts *M*(*SD*)	Low experts *M*(*SD*)
"Troubling trio"					
Suprisingness of finding	0.79	0.43	0.16	3.00 (0.77)	3.14 (1.02)
Original sample size[a]	2.20	0.03	0.46	57.71 (54.22)	98.04 (112)
p-value of original study[b]	0.42	0.68	0.11	0.01 (0.02)	0.02 (0.02)
Study Difficulty					
Method expertise required	0.64	0.52	0.13	2.36 (1.31)	2.19 (1.18)
Extant study requires diligence	0.65	0.52	0.13	2.29 (1.08)	2.15 (0.98)
Difficulty to implement	1.05	0.30	0.21	3.89 (1.35)	4.19 (1.50)
Effect robustness (original effect size)	2.21	0.03	0.43	0.44 (0.19)	0.36 (0.18)
Generalizability (context sensitivity rating from Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016a, 2016b)	1.66	0.10	0.33	2.65 (1.15)	3.04 (1.19)

Notes. All data retrieved from OSC report except generalizability. All OSC variables were rated by independent coders except implementation difficulty (rated by replication teams).
[a] One sample size value was removed because it was an outlier ($n = 230{,}047$).
[b] Two *p*-values were removed that were highly non-significant (p's = 0.91 and 0.48).

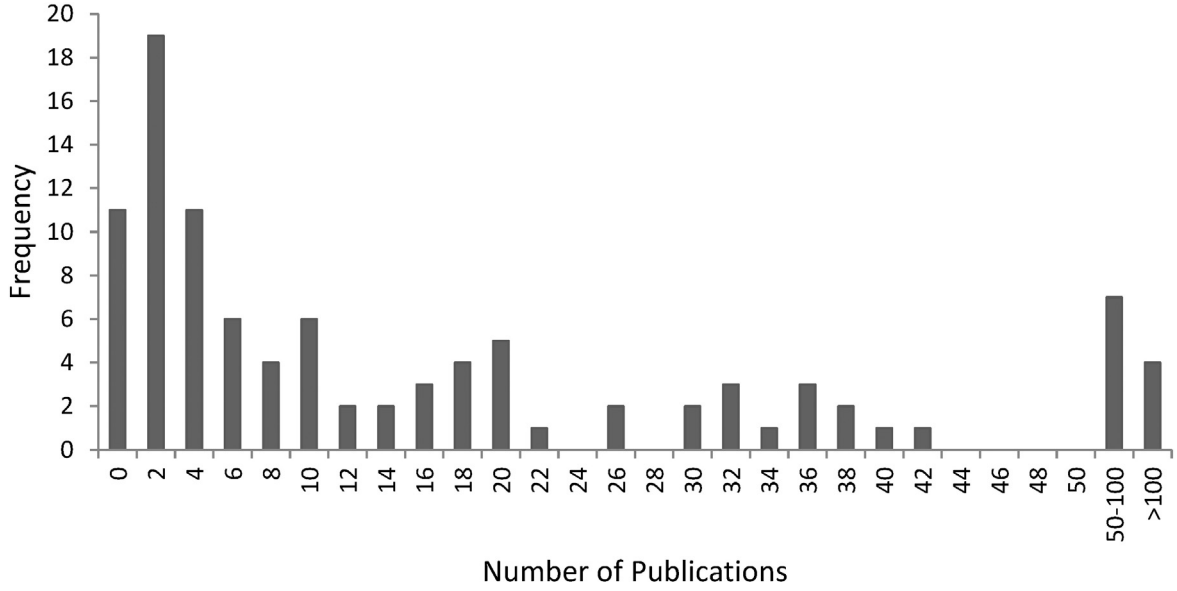

Fig. 1. Frequency distribution of combined number of publications of replication team. *Note.* The distribution of number of publications was highly positively skewed, with a large cluster of scores at or near zero. This extremely non-normal distribution affects all statistics that rely on mean values and standard deviations (because these descriptive statistics will be impacted by extreme scores and thus not be representative of the typical score) or even rank order scores (as many scores would have the identical rank and therefore the rank value for number of publications will not be informative about the relationship between number of publications and effect size).

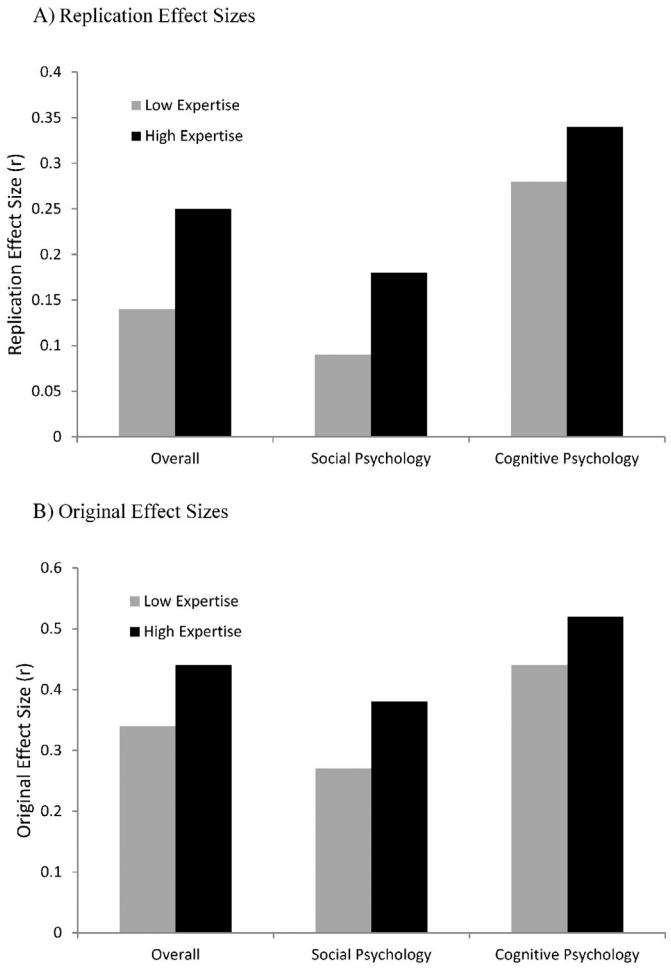

Fig. 2. Replication teams with high expertise (over 10 publications) obtained larger replication effect sizes and selected studies with larger original effect sizes than teams with low expertise.

predicted original effect size (p's < 0.001; see Table 2 in Supplemental materials).

As shown in Fig. 2, a mixed effects analysis revealed that the original effect sizes in studies selected by high-expertise replication teams were larger ($r = 0.44$, 95% CI [0.372, 0.499], $Z = 11.76$, $p < 0.001$) than those selected by low-expertise teams ($r = 0.34$, 95% CI [0.278, 0.398], $Z = 10.20$, $p < 0.001$), $Q_B(1) = 4.83$, $p = 0.028$. This difference remained after controlling for self-reported domain expertise and self-reported methodological expertise (see Table 3 in Supplemental materials). Mirroring the replication effect size results, the impact of replication team expertise was more pronounced for social psychology studies ($r_{highexpertise} = 0.38$, 95% CI [0.298, 0.458]; $r_{lowexpertise} = 0.27$, 95% CI [0.202, 0.336]; $Q_B(1) = 4.23$, $p = 0.040$) relative to cognitive psychology studies ($r_{highexpertise} = 0.52$, 95% CI [0.428, 0.606], $r_{lowexpertise} = 0.44$, 95% CI [0.354, 0.522]; $Q_B(1) = 1.65$, $p = 0.200$).

2.2.2. Expertise beyond study selection?

Given that expertise was related to larger effect sizes in original and replication studies, we conducted further analyses to explore whether the differences in replication effect sizes were significant after controlling for original effect size. We examined this by including original effect size as a covariate in a metaregression that predicted replication effect sizes from expertise and original effect size. This revealed that none of the relationships between expertise and replication effect size remained significant after controlling for original effect size (see Table 1 in Supplemental materials). Similarly, analyses that used the dichotomous variable for expertise (high, low) were not significant after controlling for original effect size (e.g., expertise based on number of publications, $b = 0.0004$, $Z = 0.98$, $p = 0.9843$). This suggests that the difference in replication success between high and low expertise teams was the result of experts selecting studies with larger original effect sizes.

2.2.3. Differences in study selection?

Given that there were differences between high and low expert replication teams in the studies that they selected, what were experts looking for in the studies? Perhaps either high or low expertise teams had a "replication axe to grind" and were thus more likely to pick studies they thought were *unlikely* to replicate. Conversely, replication teams may have differed in their motivation to select studies that

were *more likely* to replicate (e.g., perhaps more seasoned researchers have more invested in the "system" and were motivated to demonstrate the replicability of the science). Though there were no direct self-report measures of motivation, we examined several proxy variables in the OSC data set. Specifically, we examined whether replicators were differentially looking for potential "red flags" such as the troubling trio of a surprising result, small sample size, and *p*-values just under 0.05 (Lindsay, 2015). The only difference between replication teams that emerged suggested that, on average, high experts picked studies with smaller sample sizes (Table 1). However, given that similar patterns were not found for the other two aspects of the trio, there is not compelling evidence that either group was more or less motivated to select studies with commonly recognized "red flags".

A second possibility is that teams chose studies with differential levels of difficulty. Perhaps high experts were more likely to select easy studies to implement and these studies had larger original effect sizes. There was no indication of this possibility in the data.

Finally, it is also possible that replication teams relied on more subtle cues in the methodology when selecting studies to replicate. Van Bavel et al. (2016a) recently coded a "context sensitivity" variable for the studies included in the OSC (2015) report, based on "the extent to which the research topic in each study was 'contextually sensitive' (varying in time, culture or location)" (p. 6454). As reported in Table 1, there was a tendency, associated with a small to moderate effect size, for high expertise teams to select studies with more generalizable findings than low expertise teams. A recent critique of the Van Bavel paper (Inbar, in press) suggested that context sensitivity no longer predicted replication success after controlling for study domain. When domain was entered into an ANOVA with expertise, there was a main effect of domain, $F(1, 93) = 79.31$, $p < 0.001$, $\eta_p^2 = 0.460$, such that social studies were rated as more context sensitive ($M = 3.57$, $SD = 0.95$) than cognitive studies ($M = 1.96$, $SD = 0.79$), and no interaction between domain and expertise, $F(1, 93) = 0.43$, $p = 0.513$, $\eta_p^2 = 0.005$. The marginal main effect of expertise remained after controlling for domain, $F(1, 93) = 3.14$, $p = 0.080$, $\eta_p^2 = 0.033$, such that low expertise teams chose more context sensitive studies ($M = 3.04$, $SD = 1.19$) than high expertise teams ($M = 2.65$, $SD = 1.16$).

3. Discussion

The current findings support the claim that experience influences many stages of science, including the replication stage. Specifically, our findings reveal that RP:P replication teams with greater expertise selected studies to replicate that were more robust (i.e., had larger effect sizes) and generalizable (i.e., less sensitive to context) than teams with less expertise. High expertise teams also selected studies with smaller sample sizes, but did not differ from low expertise teams on selection of studies based on *p*-value or the counterintuitiveness of the finding, suggesting that this selection difference was not driven by a desire to replicate (or avoid replicating) studies with red flags.

We doubt that more seasoned researchers were literally looking for large effect sizes during the selection stage of the RP:P, but rather were more easily able to deduce which studies were robust and generalizable. Van Bavel et al. (2016a) recently found that context sensitivity was a reliable predictor of replication success (see also Inbar, in press and Van Bavel et al., 2016b for further discussion of this issue) and inferred that the experienced researchers who coded context sensitivity were able to identify characteristics that influence reproducibility. Similarly, we suggest that experienced researchers on replication teams are able to identify characteristics that predict reproducibility. The ability to identify studies with these qualities likely reflects years of training and experience that permits researchers to weigh various methodological choices. Future research should identify what methodological features experts attend to when selecting studies, and how these characteristics contribute to replication success.

While our analyses do not suggest expertise contributed to replication success above and beyond the selection effect, some caution should be taken with this interpretation. The available sample of studies was small in the OSC data and original and replication effect size were strongly correlated ($r = 0.61$). It is possible that expertise would contribute to replication effect sizes beyond selection with a larger sample of studies. Nonetheless, the current results provocatively suggest that expertise may not play a role in the execution of replication studies. Rather, expertise contributes to the ability to identify studies that are good candidates for replication.

Of course, the RPP is to date a unique effort in that researchers were selecting studies from a limited pool of preselected studies (see OSC, 2015). This is unlike other replication efforts in which researchers self-select studies from the entire vast literature. The processes that influence selection of a study to replicate in that context may differ from the RPP context. The RP:P also differs from other large scale replication efforts such as the Many Labs projects (Ebersole et al., 2015; Klein et al., 2014), in which many different researchers replicate the same set of studies. What motivates a researcher to engage in replication efforts in the first place may also differ across these various contexts and result in different types of selection issues. Thus, ironically, these results are themselves ripe for replication in order to truly understand the value of expertise in scientific research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jesp.2016.07.003.

References

Biostat (2014). *Comprehensive meta-analysis, version 3 [Computer software]* (Retrieved from) http://www.meta-analysis.com

Bornstein, M., & Cooper, H. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 221–235) (Russel Sage Foundation).

Carey, B. (2015, August 27). Many psychology findings not as strong as claimed, study says. *The New York times* (Retrieved from) http://www.nytimes.com

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Adams, R. B., Allen, J., ... Nosek, B. A. (2015). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*.

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

Inbar, Y. (2016). The association between "contextual dependence" and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences of the United States of America* (In Press).

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*, 515.

Klein, R., Ratliff, K., Vianello, M., Adams, R., Jr., Bahník, S., Bernstein, M., ... Nosek, B. (2014). Data from investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*, 142–152.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*, 1827–1832.

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

Schimmack, U. (2015). September 3). *The reproducibility of social psychology in the OSF-reproducibility project* (Retrieved November 20, 2015 from) https://replicationindex.wordpress.com/2015/09/03/comparison-of-php-curve-predictions-and-outcomes-in-the-osf-reproducibility-project-social-psychology-part-1/

Smith, E. R. (2000). Research design. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social psychology* (pp. 17–39). Cambridge: Cambridge University Press.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016a). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 6454–6459.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016b). Reply to Inbar: Contextual sensitivity helps explain the reproducibility gap between social and cognitive psychology. *Proceedings of the National Academy of Sciences of the United States of America* (In Press).

Wilson, T. D., Aronson, E., & Carlsmith, K. (2010). The art of laboratory experimentation. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology, Vol 1* (pp. 51–81) (5th ed.). Hoboken, NJ: John Wiley & Sons.