Journal of Experimental Social Psychology 69 (2017) 184–186

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Short Communication

Observe, hypothesize, test, repeat: Luttrell, Petty and Xu (2017) demonstrate good science☆

Charles R. Ebersole [a,*], Ravin Alaei [b], Olivia E. Atherton [c], Michael J. Bernstein [d], Mitch Brown [e], Christopher R. Chartier [f], Lisa Y. Chung [g], Anthony D. Hermann [h], Jennifer A. Joy-Gaba [g], Marsha J. Line [a], Nicholas O. Rule [b], Donald F. Sacco [e], Leigh Ann Vaughn [i], Brian A. Nosek [j]

[a] *University of Virginia, United States*
[b] *University of Toronto, Canada*
[c] *University of California, Davis, United States*
[d] *Pennsylvania State University, Abington, United States*
[e] *The University of Southern Mississippi, United States*
[f] *Ashland University, United States*
[g] *Virginia Commonwealth University, United States*
[h] *Bradley University, United States*
[i] *Ithaca College, United States*
[j] *Center for Open Science and University of Virginia, United States*

ARTICLE INFO

Article history:
Received 18 November 2016
Revised 1 December 2016
Accepted 5 December 2016
Available online 20 December 2016

ABSTRACT

Many Labs 3 (Ebersole et al., 2016) failed to replicate a classic finding from the Elaboration Likelihood Model of persuasion (Cacioppo, Petty, & Morris, 1983; Study 1). Petty and Cacioppo (2016) noted possible limitations of the Many Labs 3 replication (Ebersole et al., 2016) based on the cumulative literature. Luttrell, Petty, and Xu (2017) subjected some of those possible limitations to empirical test. They observed that a revised protocol obtained evidence consistent with the original finding that the Many Labs 3 protocol did not. This observe-hypothesize-test sequence is a model for scientific inquiry and critique. To test whether these results advance replicability and knowledge transfer, we conducted direct replications of Luttrell et al. in nine locations (Total $N = 1219$). We successfully replicated the interaction of need for cognition and argument quality on persuasion using Luttrell et al.'s optimal design (albeit with a much smaller effect size; $p < 0.001$; $f^2 = 0.025$, 95%CI [0.006, 0.056]) but failed to replicate the interaction that indicated that Luttrell et al.'s optimal protocol performed better than the Many Labs 3 protocol ($p = 0.135$, *pseudo* $R^2 = 0.002$). Neither Luttrell et al.'s effect size estimate for the need for cognition by argument quality interaction nor their estimate for the interaction with replication protocol fell within our corresponding 95% confidence intervals. Nevertheless, pragmatically, we favor the Luttrell et al. protocol with large samples for future research using this paradigm.

In Many Labs 3 (ML3), Ebersole et al. (2016) selected 10 original studies for replication and used 20 samples to evaluate variation in effect magnitudes in student samples across the academic semester. ML3 organizers selected Study 1 from Cacioppo, Petty, and Morris (1983, hereafter "CPM") as a "sure bet" because it represents part of a robust literature of empirical evidence for the Elaboration Likelihood Model and because it could plausibly show variability over the course of the academic semester. Surprisingly, the key CPM finding ($N = 114$, $f^2 = 0.20$, 95%CI [0.06, 0.41]) did not replicate in the ML3 samples ($N = 2365$, $f^2 < 0.001$, 95%CI [0, 0.002]).

Petty and Cacioppo (2016, hereafter "PC") offered some hypotheses for why the ML3 result differed from CPM's. Luttrell, Petty, and Xu (2017, hereafter "LPX") put some of those hypotheses to empirical test. They revised the ML3 protocol in some ways to make it more similar to CPM and incorporated insights from other research that differed from CPM but might maximize the effect. Participants randomly assigned to the ML3 protocol did not show evidence for the original finding ($N = 106$, $p = 0.60$, $f^2 = 0.001$, 95%CI [0, 0.057]), but participants randomly assigned to LPX's revised protocol did show evidence for the original finding ($N = 108$, $p = 0.01$, $f^2 = 0.07$, 95%CI [0.003, 0.196]). The key result was that the interaction between need for cognition and argument quality on persuasion was larger in the optimized

☆ We would like to thank Andrew Luttrell for sharing materials for this study. CE and BN wrote the report. CE and ML conducted all analyses. All authors contributed to data collection and revising the report.

DOI of original article: http://dx.doi.org/10.1016/j.jesp.2016.09.006.

* Corresponding author.
E-mail address: cebersole@virginia.edu (C.R. Ebersole).

http://dx.doi.org/10.1016/j.jesp.2016.12.005
0022-1031/© 2016 Elsevier Inc. All rights reserved.

Table 1
Descriptive statistics and summary of key effects for each collection site.

Site	N	% Female	M Age	SD Age	NFC × AQ × Replication Type			LPX optimal: NFC × AQ			ML3: NFC × AQ		
					t-Value	*p*-Value	η_p^2	t-Value	*p*-Value	η_p^2	t-Value	*p*-Value	η_p^2
University of Virginia	157	57.3	18.68	1.00	1.74	0.084	0.021	2.63	0.011	0.091	0.02	0.985	<0.001
Ashland University	117	70.1	19.4	2.86	0.71	0.478	0.005	1.83	0.074	0.058	0.54	0.595	0.005
Bradley University	181	69.6	18.9	1.47	−0.15	0.885	<0.001	0.98	0.332	0.011	0.89	0.377	0.009
Ithaca College	72	84.7	19.6	1.18	0.07	0.944	<0.001	0.13	0.901	0.001	−0.12	0.907	<0.001
Pennsylvania State University-Abington	174	71.3	19.7	2.80	0.36	0.721	0.001	1.00	0.320	0.012	0.40	0.688	0.002
University of California-Davis	189	75.1	20.1	2.29	0.72	0.472	0.003	0.49	0.627	0.003	−0.54	0.593	0.003
University of Toronto	178	80.3	18.7	2.80	0.98	0.328	0.006	2.52	0.014	0.070	1.41	0.162	0.023
University of Southern Mississippi	139	86.3	21.3	5.48	0.22	0.823	<0.001	0.05	0.965	0.000	−0.27	0.788	0.001
Virginia Commonwealth University	20	80.0	19.7	1.45	0.22	0.828	0.004	0.40	0.705	0.031	0.20	0.851	0.008

LPX protocol compared to the ML3 protocol ($p = 0.03$, $f^2 = 0.02$, 95%CI [0, 0.081]).

LPX provided information about which factors may relate to eliciting and detecting the original effect. PC pointed out that the arguments in ML3 were possibly too short (approximately 165 words, compared to approximately 300-word arguments in CPM), so LPX used much longer arguments (~900 words) than those used in either ML3 or the original CPM. PC also suggested that ML3's weak arguments were not sufficiently weak. However, LPX's weak arguments ($M = 5.49$ on a 9-point scale, $SD = 1.66$) were descriptively rated as stronger than ML3's ($M = 5.29$, $SD = 1.58$). PC also argued that the key effect is most detectable when the presented arguments do not have high personal relevance. This was not part of the original CPM but LPX explicitly stated that the topic of the arguments, the introduction of comprehensive exams for undergraduate seniors, would not affect the participants. Finally, PC suggested that the use of a shortened Need for Cognition (NFC) scale might have reduced effect detectability. However, the key effect in LPX was statistically reliable, whether using LPX's 18-item scale or just the five items of that scale used in ML3.

Descriptively, these results suggest that some of PC's hypotheses have merit for observing the persuasion effect whereas others may not. The sequence of ML3's evidence, PC's hypothesizing, and LPX's testing is a model for investigating the replicability of research and for advancing theoretical understanding of observed outcomes (Klein et al., 2014b). An initial replication attempt (ML3) generated hypotheses about which methodological features were necessary to observe an effect (PC). A new investigation (LPX) provided support for some of these features but not others. As a next step in this iterative process, we sought to independently validate LPX's findings, testing whether the expertise provided by LPX's design could be successfully replicated in a large-sample preregistered design by independent researchers.

To achieve this, ML3's original contributors were invited to participate in a crowdsourced replication of LPX, including random assignment to test the comparative effectiveness of the ML3 and LPX protocols. We strived to collect as many participants as possible before the end of the academic term and did not analyze any data until the end of collection. In total, nine sites contributed 1219 participants. The same study script from LPX was used, revising only the year referenced and the name of the university to match the current year and location of each collection site. The analysis plan was preregistered on the Open Science Framework and is available at: https://osf.io/chxja/. Furthermore, this introduction was drafted before the results of this replication were known (but revised later for clarity and style). Details of each sample and data collection site are presented in Table 1, and all data, materials, and supplementary analyses are available at https://osf.io/x96at/.

1. Results

LPX's main claim was that their optimized protocol provided a significant improvement over the ML3 protocol in terms of detecting the focal Need for Cognition (NFC) × Argument Quality (AQ) interaction that predicted persuasion in CPM. A total of 1274 participants provided at least one response; 1219 provided all needed responses to be included in the analyses. With this sample size, we had 99.9% power to detect LPX's observed effect size of $f^2 = 0.02$ for the key 3-way interaction, 95% power to detect an effect size of $f^2 = 0.011$, and 80% power to detect an effect size of $f^2 = 0.006$.

To test the key claim in our replications, we submitted the data to a hierarchical mixed-effects model. Step 1 contained initial attitudes toward comprehensive exams, AQ, Replication Type, and NFC as simultaneous fixed effects predictors of message evaluation with collection site as a random intercept. Step 2 added all corresponding two-way interactions as fixed effects. Step 3 added the focal three-way interaction of NFC × AQ × Replication Type. The addition of the three-way interaction did not significantly improve the model, X^2 (1, $N = 1219$) $= 2.23$, $p = 0.135$, *pseudo* $R^2 = 0.002$. That is, LPX's protocol did not provide a significant improvement over the ML3 protocol in these data. Furthermore, collapsing across collection site to increase comparability with LPX's model, our confidence interval for the three-way interaction does not contain LPX's estimate ($f^2 = 0.02$), $b = 0.11$, $SE = 0.07$, $t(1210) = 1.49$, $p = 0.137$, $f^2 = 0.002$, 95% CI [0, 0.010].[1]

The overall model did, however, show a reliable interaction of NFC and AQ predicting message evaluation, replicating the original effect, $b = 0.27$, $SE = 0.07$, $t(1206) = 3.75$, $p < 0.001$. Although the overall model did not provide evidence for moderation by replication type, we next examined the original NFC × AQ interaction within each of the LPX and ML3 replications. Collapsing across collection sites and retaining initial attitudes as a covariate like LPX did, NFC and AQ significantly interacted to predict message evaluation using the LPX protocol, $b = 0.39$, $SE = 0.10$, $t(602) = 3.86$, $p < 0.001$, $f^2 = 0.025$, 95% CI [0.006, 0.056]. The same interaction did not emerge under the ML3 protocol, however, $b = 0.16$, $SE = 0.10$, $t(607) = 1.57$, $p = 0.117$, $f^2 = 0.004$, 95% CI [0, 0.020]. Of note, LPX's effect size estimate for the NFC × AQ interaction ($f^2 = 0.07$) does not fall within our 95% confidence interval.

2. Discussion

With high-power to detect LPX's effects, we replicated some of their results but not others. Unlike ML3, we obtained evidence for the critical NFC × AQ interaction found by both CPM and LPX, though with a much weaker effect size. In comparing the LPX and ML3 protocols, we found that the LPX version returned a significant effect but the ML3 version did not. However, the key three-way interaction testing whether the protocols reliably differed was not significant. As it is inappropriate to interpret two effects as different simply because one's significance

[1] We primarily compare LPX's effect sizes to our confidence intervals, rather than the reverse, because of our intervals' greater precision. Our estimates of the key three-way interaction and the NFC × AQ interaction using the LPX protocol fall within the corresponding 95% confidence intervals from LPX. However, the lower bounds of those intervals are $f^2 = 0$ and $f^2 = 0.003$, respectively, meaning that a failure to replicate by that metric would require an effect of very nearly 0 or a reversal of the effect.

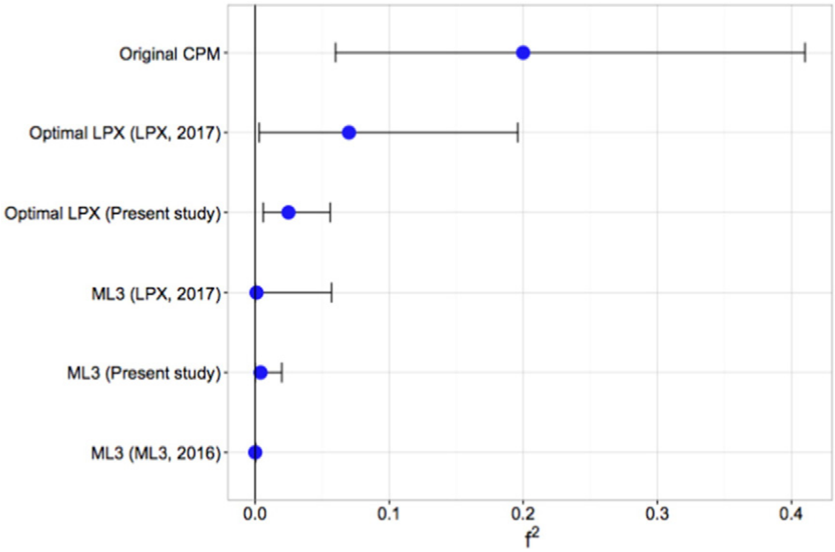

Fig. 1. Effect sizes and 95% CIs for experiments using the original CPM protocol, optimized LPX protocol, and ML3 protocol. *Note.* CPM = Cacioppo, Petty, and Morris (1983); LPX = Luttrell, Petty, and Xu (2017); ML3 = Ebersole et al. (2016). The minimum possible value for CI lower bound is 0. CI for ML3 (2016) is not visible because it is within the size of the effect size bullet [0, 0.002]. Cohen (1988) suggested the following benchmarks for interpreting f^2 effect sizes: 0.02 for a small effect, 0.15 for a medium effect, and 0.35 for a large effect.

level falls below the $p = 0.05$ threshold and the other's does not, we instead rely on the nonsignificant difference between our comparison of the two versions but caution that our study was perhaps underpowered to detect a difference between them. A sample of 6500 is needed for 95% power to detect the 3-way interaction effect size we observed. Thus, based on the available evidence, we would recommend that a researcher selecting a protocol to study variations in NFC × AQ effects on persuasion use the LPX protocol.

Fig. 1 shows the observed effect sizes and confidence intervals of the key interaction of NFC × AQ on persuasion from the original CPM, from our first large-scale replication attempt in ML3, from LPX's comparison of ML3 to their optimized version, and from our present large-scale replication of the LPX and ML3 comparison. Two findings stand out. Our large sample and preregistered replications produced estimates that are weaker and more precise. Neither CPM's nor LPX's optimized protocol effect sizes fall within the confidence interval of any of our replications, including our replication of LPX's optimized protocol. This is particularly surprising, given that the protocol is easily adapted and similarly relevant for undergraduate students at the tested institutions. Further, the differences are unlikely to be attributable to experimenter effects, quality of design, or execution because we used the same materials as LPX and data collection was automated. It is also notable that the original CPM effect far exceeds the confidence interval of our high-powered replication of LPX's optimized design, and even exceeds the relatively wide confidence interval of the LPX data collection with the optimized design. The original study effect size is an outlier compared to all other versions and data collections.

Based on the present evidence, we conclude that the NFC × AQ effect on persuasion in this paradigm is reliable, but also up to 88% weaker than originally observed by CPM, and 64% weaker than observed in LPX's initial test of their optimized design. Based on the effect size we observed, effective study of this phenomenon using LPX's optimized protocol requires sample sizes of 316 for 80% power and 522 for 95% power.

Accumulating evidence suggests that reproducibility of evidence in psychology is more challenging than expected or desired (e.g., Ebersole et al., 2016; Klein et al., 2014a; Open Science Collaboration, 2015). This has elicited a variety of reactions in response to failures to replicate. In this case, PC and LPX generated hypotheses to explain differences between CPM and ML3, and then conducted an investigation generating independent data to test those hypotheses. With this observe-hypothesize-test sequence, PC and LPX treated the different outcomes of CPM and ML3 as worthy of study rather than simply hypothesizing about the failure to replicate in defense of the original results. In this regard, Luttrell, Petty, and Xu have provided a model of productive scientific critique worth emulating.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jesp.2016.12.005.

References

Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of Personality and Social Psychology*, *45*(4), 805–818.

Cohen, J. E. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014a). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, *45*, 142–152. http://dx.doi.org/10.1027/1864-9335/a000178.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014b). Theory building through replication: Response to commentaries on the “Many Labs” replication project. *Social Psychology*, *45*, 307–310.

Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology*, *69*, 178–183.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. http://dx.doi.org/10.1126/science.aac4716.

Petty, R. E., & Cacioppo, J. T. (2016). Methodological choices have predictable consequences in replicating studies on motivation to think: Commentary on Ebersole et al. (2016). *Journal of Experimental Social Psychology*, *67*, 86–87.