Journal of Experimental Social Psychology 69 (2017) 178–183

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Case Report

Replicating and fixing failed replications: The case of need for cognition and argument quality

Andrew Luttrell [a],*, Richard E. Petty [b],**, Mengran Xu [b]

[a] *College of Wooster, United States*
[b] *The Ohio State University, United States*

ARTICLE INFO

Article history:
Received 6 June 2016
Revised 12 September 2016
Accepted 12 September 2016

Keywords:
Replication
Reproducibility
Need for cognition
Elaboration likelihood model

ABSTRACT

Recent large-scale replication efforts have raised the question: how are we to interpret failures to replicate? Many have responded by pointing out conceptual or methodological discrepancies between the original and replication studies as potential explanations for divergent results as well as emphasizing the importance of contextual moderators. To illustrate the importance of accounting for discrepancies between original and replication studies as well as moderators, we turn to a recent example of a failed replication effort. Previous research has shown that individual differences in need for cognition interact with a message's argument quality to affect evaluation (Cacioppo, Petty, & Morris, 1983). However, a recent attempt failed to replicate this outcome (Ebersole et al., 2016). We propose that the latter study's null result was due to conducting a non-optimal replication attempt. We thus conducted a new study that manipulated the key features that we propose created non-optimal conditions in the replication effort. The current results replicated the original need for cognition × argument quality interaction but only under the "optimal" conditions (closer to the original study's method and accounting for subsequently identified moderators). Under the non-optimal conditions, mirroring those used by Ebersole et al., results replicated the failure to replicate the target interaction. These findings emphasize the importance of *informed replication*, an approach to replication that pays close attention to ongoing developments identified in an effect's broader literature.

1. Introduction

As any comment on replication must acknowledge, reproducibility is integral to the scientific enterprise. Recently, however, several large-scale efforts to replicate previous findings in psychology have claimed failure to find evidence for many of the original effects (e.g., Ebersole et al., 2016; Klein et al., 2014; Open Science Collaboration, 2015).

One response to such failures is to highlight elements that differed between the original and replication studies. For example, Gilbert, King, Pettigrew, and Wilson (2016) suggested that the materials used in some prominent replication attempts (e.g., Open Science Collaboration, 2015) were not very faithful to those of the original studies and that these discrepancies were associated with replication failure. Being faithful to the original study, however, can be defined in at least three ways. First, a replication could be criticized for failing to *exactly* replicate the original study, omitting or modifying critical elements in the methodology (cf. Brandt et al., 2014; Simons, 2014). Second, a replication could be criticized for failing to *conceptually* replicate the study (e.g., Crandall & Sherman, 2016; Fabrigar & Wegener, 2016). That is, sometimes adhering too strictly to original materials and procedures may fail to capture the key psychological concepts of interest in a new sample or setting. Third, replication efforts can also fail to account for theoretically relevant moderators even if concepts are operationalized appropriately. The original effect may not be false—it just occurs under particular conditions (e.g., Cesario, 2014; Dijksterhuis, 2014; Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016).

Typically, this is where discussions of replication failures end. Rarely, if ever, is a new study conducted to show that a replication will be successful if it employs optimal procedures but will fail if it uses the non-optimal procedures for which a failed replication study was criticized. Indeed, some have argued that criticisms of replication studies are mostly post-hoc and speculative, noting that such critiques instead present testable claims and that researchers should conduct a study "to demonstrate that they can reproduce the effect and make it vanish" (Simons, 2014, p. 77). We aim to do just that.

The effect in question is the interaction between individuals' enjoyment of effortful thinking—as assessed with the need for cognition (NFC) scale—and argument quality (AQ) on the perceived convincingness

* Correspondence to: A. Luttrell, Department of Psychology, College of Wooster, Wooster, OH 44691, United States.
** Correspondence to: R.E. Petty, Department of Psychology, The Ohio State University, Columbus, OH 43210, United States.
E-mail addresses: aluttrell@wooster.edu (A. Luttrell), petty.1@osu.edu (R.E. Petty).

http://dx.doi.org/10.1016/j.jesp.2016.09.006
0022-1031/© 2016 Elsevier Inc. All rights reserved.

of a persuasive message. Cacioppo et al. (1983), hereafter called "CPM," first demonstrated that AQ (strong vs. weak) produced a larger impact on ratings of message persuasiveness for people high versus low in NFC. This finding is consistent with the Elaboration Likelihood Model (ELM; Petty & Cacioppo, 1986) and has been shown several times since the original study (for meta-analyses, see Cacioppo, Petty, Feinstein, & Jarvis, 1996; Carpenter, 2015).

A recent attempt to replicate the CPM result, part of the "Many Labs 3" project (Ebersole et al., 2016), hereafter called "ML3," failed to produce the NFC × AQ interaction, only finding a main effect of AQ. As Petty and Cacioppo (2016) noted in their comment on ML3, however, there are several discrepancies between the original study and the replication effort. Four features of ML3's materials and analysis were highlighted. First, their messages were unusually brief—about half as long as those used by CPM, and even shorter than those typically used in similar research. Second, ML3 clearly stated that the advocated proposal was targeted at participants' own universities for immediate adoption, a feature absent from CPM. Third, ML3 used an unvalidated 6-item NFC scale rather than the longer validated scales used in most prior studies. Finally, ML3 did not adequately account for a potential confound noted by CPM in which NFC was linked to initial attitudes on the senior comprehensive exams topic used (i.e., higher NFC was associated with more favorable attitudes toward the exams). To address this, CPM recruited high and low NFC participants who reported similar attitudes toward the issue in a pretest whereas ML3 did not control for initial attitudes in any way.

These differences are not trivial and plausibly contributed to the failed replication. First, because the messages used in ML3 were very short, they may have appeared quite easy to process. Research since CPM has shown that people low in NFC, who otherwise are relatively low in their motivation to think, can become more motivated to think when the information seems simple to process. In contrast, high NFC individuals become less motivated to process information when it seems simple and therefore unchallenging (See, Petty, & Evans, 2009; Wheeler, Petty, & Bizer, 2005). To the extent that these effects are an outcome of using a very brief message, the NFC × AQ interaction would be less likely to occur. Second, because the issue was made highly relevant in the ML3 replication attempt, people could be motivated to process the message carefully, regardless of their NFC (Petty & Cacioppo, 1979, 1990). Indeed, when situational variables prompt greater elaboration, NFC is no longer related to outcomes of interest in the typical way (e.g., Calanchini, Moons, & Mackie, 2016; Smith & Petty, 1996). Third, using short forms of established scales, even when informed by some empirical criteria, can pose a threat to the scale's reliability and validity, therefore making reported effects of the scales more difficult to observe (Widaman, Little, Preacher, & Sawalani, 2011). Notably, some recent research has demonstrated greater predictive ability for longer than shorter forms of the NFC and other scales (Bakker & Lelkes, 2016). Finally, without accounting for initial attitudes toward the policy, it is possible that participants low in NFC might be motivated to elaborate on the message simply because they oppose the policy more than those higher in NFC (i.e., counterattitudinal messages can provoke more processing than proattitudinal ones; cf. Cacioppo & Petty, 1979; Clark & Wegener, 2013). If so, this too would reduce the likelihood of observing the NFC × AQ interaction.

In essence, as Petty and Cacioppo (2016) argued, ML3's replication of CPM was not optimal.[1] These are only conceptual arguments, however. It remains unclear whether these factors really matter. The present study aimed to address these issues by conducting a replication of the NFC × AQ interaction under two conditions: *non-optimal*, mirroring those used by ML3, and *optimal*, accounting for the critique made by Petty and Cacioppo (2016). The Petty and Cacioppo critique considered differences in procedures between CPM and ML3 as well as developments on this topic following the original CPM publication. As such, the materials in the optimal condition of the present study do not exactly match those used in CPM but instead reflect what Petty and Cacioppo argued were the optimal conditions for finding the effect (i.e., lengthier messages, explicitly low personal relevance, a validated full NFC scale, and accounting for initial attitudes). We anticipated that the NFC × AQ interaction observed by CPM would replicate under the optimal conditions, but not under the non-optimal conditions employed by ML3.

2. Method

2.1. Participants and design

Two-hundred fourteen Ohio State University undergraduates (98 male, 115 female, 1 unreported; $M_{\text{age}} = 19.32$, $SD = 2.16$) participated in partial fulfillment of an Introductory Psychology requirement.[2] Each participant was randomly assigned to one of the four conditions comprising the 2 (Argument Quality: Weak vs. Strong) × 2 (Replication type: Optimal vs. Non-optimal procedures) between-subjects factorial design. NFC was measured.

2.2. Procedure

The study followed the basic procedure used by CPM and ML3 and was administered as an online survey. Participants first completed the NFC scale and reported their initial attitudes toward a policy that would require college seniors to take a comprehensive exam in order to graduate. They then read a message arguing in favor of the proposed policy. Half of the participants saw a message containing strong arguments whereas the other half saw a message containing weak arguments. For participants in the *non-optimal* condition, the message was relatively short and highly personally relevant (mirroring the conditions of ML3), and for participants in the *optimal* condition, the message was relatively lengthy and less personally relevant. Finally, participants reported their evaluations of the message on the scales used by both CPM and ML3. All materials are provided in the Online supplement.

2.3. Independent variables

2.3.1. Replication type: non-optimal vs. optimal procedures

In the *non-optimal* condition, the topic was made especially relevant by specifying that the senior comprehensive exam policy would be implemented immediately at the participants' university. Whereas ML3's message included this information in the message text, we also included it in the message's introduction. The messages were also relatively short (approximately 165 words) and indeed were the same messages used by ML3. In the *optimal* condition, the topic was made less relevant

[1] Petty and Cacioppo also noted other differences such as ML3's use of weak arguments that were not as specious as in CPM. This also could have influenced the failure to replicate but we do not address that here.

[2] The sample size was determined as follows. The critical NFC × AQ interaction effect size in CPM was equivalent to $f^2 = 0.20$. We submitted a more conservative effect size estimate ($f^2 = 0.10$) to an *a priori* power analysis (Faul, Erdfelder, Buchner, & Lang, 2009), setting power to 0.90 at $\alpha = 0.05$. The resulting sample size ($n = 108$) to obtain the key interaction under the optimal conditions was then doubled to account for the non-optimal conditions. In other words, we computed the sample size needed to detect the key interaction under optimal conditions and used the same sample size for the non-optimal condition to keep the number of people per cell roughly equal. A potentially better way to estimate the effect size expected for the optimal condition is to use the effect size reported in a meta-analysis. Cacioppo et al. (1996) analyzed five studies testing the NFC × AQ interaction and computed an effect size equivalent to $f^2 = 0.07$. Entering this as the expected effect size in the same power analysis shows that $N = 115$ is sufficient to achieve 0.80 power ($N = 153$ for 0.90 power), which is consistent with the sample size arrived at in our original analysis. Notably, we based these power analyses on the size of the NFC × AQ interaction because the key prediction is only that AQ will have a larger effect on message evaluation at increasing levels of NFC. This could mean, for example, that AQ will have zero effect at lower levels of NFC or just that AQ will have a smaller effect at lower than at higher levels of NFC. Thus, the NFC × AQ interaction was the focal test in this study and power analyses were conducted as such. We acknowledge, however, that other perspectives hold that power should consider specific predicted simple effects and n per cell in addition to an overall interaction (Simonsohn, 2014; Simonsohn, Nelson, & Simmons, 2014).