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/

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; $M_{age} = 19.32$, $SD = 2.16$) 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).

(i.e., the policy was said not to be implemented for another 10 years) and the messages were relatively long (approximately 900 words). These messages were adapted from Petty and Cacioppo (1986).

It is important to note that the messages used in the optimal condition were not identical to those used by CPM. First, CPM had participants read "an approximately 300-word message" (p. 812), which is about twice as long as the messages used by ML3 but still shorter than the messages used in this study. The CPM messages are no longer available and it was therefore impossible to use them.[3] Thus, in the optimal condition we chose to use the full strong and weak messages that Petty and Cacioppo reprinted in their 1986 monograph with minor modifications. These messages have been extensively pretested and have been available to the field for the past 30 years. Second, CPM did not explicitly state when the exam policy would go into effect. In the present materials, we incorporated this element as it was contained in ML3's materials. Although these aspects of the messages differed from CPM, they were included to make the condition more explicitly "optimal" for finding the NFC × AQ interaction, as outlined previously.

#### 2.3.2. Argument quality: strong vs. weak

To manipulate AQ, in both the optimal and non-optimal conditions, we adapted the gist of the commonly used strong and weak arguments in favor of senior comprehensive exams provided by Petty and Cacioppo (1986). In the non-optimal condition, the content of the messages was taken directly from ML3's materials, whereas in the optimal condition, they were taken from Petty and Cacioppo (1986) with very minor modifications mostly for readability.[4] The only substantive change was that the argument in the weak version linking failure to implement the exams to racial discrimination was changed to instead reference gender discrimination.

#### 2.3.3. Need for cognition

All participants responded to the 18-item NFC Scale (Cacioppo, Petty, & Kao, 1984).[5] ML3 only used six items to measure NFC, which they selected on the basis of item factor loadings from unpublished data. Five of the items used in ML3's reduced scale were included in the validated 18-item scale, so we also computed NFC scores based only on those five items. The 6th item used by ML3—"*More often than not, more thinking just leads to more errors*"—was a question from the original, longer version of the NFC scale (Cacioppo & Petty, 1982) that was not retained in the 18-item short measure. Thus, we could not include this item in our analyses.[6] Internal reliability was good for both the full scale ($\alpha = 0.90$) and for the 5-item reduced scale ($\alpha = 0.70$).

### 2.4. Dependent variable: message evaluation

Participants rated the perceived quality of the persuasive message using the same five items employed by both CPM and ML3 (e.g., "to what extent do you think the communication made its point effectively?" from "not at all" to "completely"). Responses were given on 9-point scales ($\alpha = 0.92$) and averaged to form an index such that higher numbers indicate a more favorable evaluation of the message.[7]

### 2.5. Covariate: initial attitudes

After reading a short general introduction to the issue but before reading the message, all participants reported their attitudes toward the exam policy using a single item measure (i.e., "To what extent would you favor such a senior exam requirement?" with responses indicated on an 11-point scale anchored at "strongly oppose" and "strongly favor"; $M = 4.57$, $SD = 2.58$). As noted above, our aim was to use this as a covariate to control for initial attitudes, mirroring the procedure used by CPM.

## 3. Results

As suggested by CPM, greater NFC was associated with more favorable pre-message attitudes toward the exam policy, $r(212) = 0.17$, $p = 0.02$. To account for this, we included initial attitudes as a covariate in the analyses, which also means that the results are more compatible with those of CPM, who specifically recruited participants so as to make NFC and initial attitudes independent.

Raw means and standard deviations of message evaluation within each experimental condition are presented in Table 1.

### 3.1. Overall three-way interaction

Data were first submitted to a hierarchical multiple regression model with message evaluation as the dependent variable. The first step of the model included initial attitudes, AQ, replication type, and NFC (18 items) as simultaneous predictors. AQ and replication type were effects coded (−1: Weak Message/Non-optimal Condition; +1: Strong Message/Optimal Condition). NFC was entered without mean-centering.[8] All corresponding two-way interaction terms were entered in the second step, and the three-way interaction term was entered in the third step (Table 2). Results are interpreted from the first step in which they appear.

First, there was a main effect of AQ such that participants reported more favorable evaluations of the message in the strong ($M = 6.18$, $SD = 1.51$) than in the weak ($M = 5.39$, $SD = 1.62$) arguments condition, $B = 0.38$, $t(209) = 3.64$, $f^2 = 0.06$, $p < 0.001$, 95% CI: [0.18, 0.59].[9] There was also a main effect of NFC such that higher NFC was associated with less favorable post-message evaluations, $B = -0.36$, $t(209) = -2.15$, $p = 0.03$, $f^2 = 0.02$, 95% CI: [−0.69, −0.03]. Finally, there was a main effect of initial attitudes such that more positive initial attitudes correspond to more positive evaluations of the pro-policy message, $B = 0.12$, $t(209) = 2.89$, $f^2 = 0.04$, $p = 0.004$, 95% CI: [0.04, 0.20].

Most importantly, the data supported the expected three-way interaction, $B = 0.37$, $t(205) = 2.20$, $f^2 = 0.02$, $p = 0.03$, 95% CI: [0.04, 0.70] (see Fig. 1).[10] This effect was such that the NFC × AQ interaction was not

[3] The original publication did not provide the full text of the messages nor are we aware of any records that remain from that study, but CPM noted that that the set of arguments used "were essentially those described as the 'strong' and 'very weak' communications in [Petty, Harkins, and Williams] (1980)" but also emphasized that those original messages from Petty, Harkins, and Williams (1980) had been further altered based on pre-testing (p. 808). Nevertheless, Petty et al. (1980) did not provide the full messages either, but they reported a summary of the arguments contained in the messages used (see Supplementary materials), which are somewhat different from the messages used in the optimal condition and by CPM.

[4] For example, the original messages included the phrase "At comparable schools without the exams…" The messages used in the present study merely clarified this, instead writing, "At comparable schools that did not implement the exams…"

[5] CPM used the original 34-item NFC scale (Cacioppo & Petty, 1982); however, the shorter 18-item scale has since become the standard measure.

[6] We only became aware of this issue after data collection. In their article, ML3 did not report the specific items used in their 6-item scale, so we initially modeled their selection process and chose as the first six items of the scale those statements which showed the highest correlations with total NFC scores and highest factor loadings in the scale's first published report (Cacioppo & Petty, 1982, Study 1). These were presented on one page of the online survey and the remaining 12 items were presented on the next page. Our original approach was to use these six items ($\alpha = 0.84$) to test the impact of using reduced scales. Upon the editor's request, we contacted Ebersole et al. to obtain the specific items they used and discovered that one of the items was not included in the scale we measured. We thus present results with a 5-item reduced scale to most closely approximate the exact scale used in ML3. See Supplementary materials for analyses with the originally intended 6-item reduced scale.

[7] Like ML3, some work following CPM has continued to use these items to test for the NFC × AQ interaction (see Cacioppo et al., 1996), but other work has used a measure of attitudes toward the issue (e.g., senior comprehensive exams) instead and looked for the same interaction (see Carpenter, 2015, for a review).

[8] The choice not to mean-center NFC scores followed from recommendations by Hayes (2013). Although it is common to mean-center variables in multiple regression analyses, it does not affect the significance level of the target effects.

[9] Reported means are the raw means and all confidence intervals are around the regression coefficients ($B$), unless otherwise noted.

[10] When not controlling for initial attitudes, the three-way interaction ($p = 0.06$) is somewhat weaker, but the NFC × AQ interaction still emerges under the optimal conditions ($p = 0.02$), but not the non-optimal conditions ($p = 0.78$). See Supplementary analyses for a full report.

**Table 1**
Mean and standard deviations of message evaluation by argument quality and replication type.

| | Weak arguments | | Strong arguments | |
|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* |
| Non-optimal condition | 5.29 | 1.58 | 5.88 | 1.66 |
| Optimal condition | 5.49 | 1.66 | 6.46 | 1.31 |

significant in the non-optimal (replication of ML3) condition, $B = -0.12$, $t(205) = -0.53$, $f^2 = 0.001$, $p = 0.60$, 95% CI: [−0.56, 0.33], but the same interaction was significant in the optimal condition, $B = 0.62$, $t(205) = 2.50$, $f^2 = 0.03$, $p = 0.01$, 95% CI: [0.13, 1.11]. Within the non-optimal condition, there was only an effect of AQ, estimated at the mean value of NFC, $B = 0.30$, $t(205) = 2.02$, $f^2 = 0.02$, $p = 0.05$, 95% CI: [0.01, 0.60], with strong arguments producing more positive evaluations than weak (see bottom panel of Fig. 1). Within the optimal condition, the significant NFC × AQ interaction was such that there was no AQ effect at relatively low (1 SD below the mean) NFC, $B = 0.14$, $t(205) = 0.71$, $p = 0.48$, 95% CI: [−0.25, 0.53], but strong arguments led to more favorable evaluations than weak arguments at relatively high (1 SD above the mean) NFC, $B = 0.93$, $t(205) = 4.00$, $p < 0.001$, 95% CI: [0.47, 1.39], replicating the result reported by CPM (see top panel of Fig. 1). As in the non-optimal condition, there was also an effect of AQ, estimated at the mean value of NFC, $B = 0.54$, $t(205) = 3.63$, $f^2 = 0.06$, $p < 0.001$, 95% CI: [0.25, 0.83], again showing that strong arguments led to more favorable evaluations than weak.

### 3.2. Analyses within the optimal condition only

In the previous analysis, the NFC × AQ interaction in the optimal condition was tested in the context of the full regression model examining the three-way interaction. That is, the effect was estimated in a regression model that set replication type at "optimal" (+1) but that still controlled for the other two-way interaction terms and the three-way interaction term. Because the effect size ($f^2$) is based on the unique variance explained by the regression term while also accounting for the variance explained by the whole regression model (Cohen, 1988), the effect size is bound to be smaller in a model with other potentially related terms in it. In other words, by including replication type and the other interaction terms, we might have underestimated the size of the NFC × AQ effect under the optimal conditions.

Thus, we report a new analysis conducted on data from the optimal condition only ($n = 108$). This would be commensurate to a focused replication of the NFC × AQ effect under optimal conditions. Data were submitted to a hierarchical multiple regression model predicting message evaluation, entering initial attitudes, NFC, and AQ in Step 1, and NFC × AQ in Step 2. Results are interpreted from the first step of the model in which they appear. First, a marginal main effect of initial attitudes appeared such that more positive initial attitudes toward the policy were associated with more positive evaluations of the message, $B = 0.10$, $t(104) = 1.81$, $f^2 = 0.03$, $p = 0.07$, 95% CI: [−0.01, 0.21]. There was also a main effect of NFC such that higher NFC was associated with less positive evaluations of the message, $B = -0.51$, $t(104) = -2.02$, $f^2 = 0.04$, $p = 0.05$, 95% CI: [−1.00, −0.01]. Finally, there was a main effect of AQ such that message evaluations were more positive in the strong arguments (vs. weak arguments) condition, $B = 0.48$, $t(104) = 3.38$, $f^2 = 0.11$, $p = 0.001$, 95% CI: [0.20, 0.76].

Most importantly, there was a significant NFC × AQ interaction, $B = 0.62$, $t(103) = 2.61$, $f^2 = 0.07$, $p = 0.01$, 95% CI: [0.15, 1.08]. The interaction was such that there was no AQ effect at relatively low levels of NFC (1 SD below the mean), $B = 0.12$, $t(103) = 0.60$, $p = 0.55$, 95% CI: [−0.27, 0.51], but at relatively high levels of NFC (1 SD above the mean), strong arguments led to more positive evaluations of the message than weak arguments, $B = 0.84$, $t(103) = 4.30$, $p < 0.001$, 95% CI: [0.45, 1.23]. This analysis within just the optimal conditions produces an effect size for the key NFC × AQ interaction ($f^2 = 0.07$) that is roughly twice the size estimated from the earlier analysis controlling for all conditions and interactions ($f^2 = 0.03$).

**Table 2**
Multiple regression models predicting message evaluation.

| | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|
| | *B* | *p* | *B* | *p* | *B* | *p* |
| Intercept | 6.41 | <0.001 | 6.55 | <0.001 | 6.45 | <0.001 |
| Initial attitude | 0.12 | 0.004 | 0.12 | 0.004 | 0.13 | 0.002 |
| Argument quality | 0.38 | <0.001 | −0.32 | 0.57 | −0.40 | 0.47 |
| Need for cognition | −0.36 | 0.03 | −0.41 | 0.02 | −0.39 | 0.02 |
| Replication type | 0.19 | 0.08 | 0.61 | 0.28 | 0.75 | 0.18 |
| AQ × NFC | | | 0.21 | 0.20 | 0.25 | 0.14 |
| AQ × replication type | | | 0.12 | 0.28 | −1.09 | 0.05 |
| NFC × replication type | | | −0.13 | 0.44 | −0.17 | 0.31 |
| AQ × NFC × replication type | | | | | 0.37 | 0.03 |

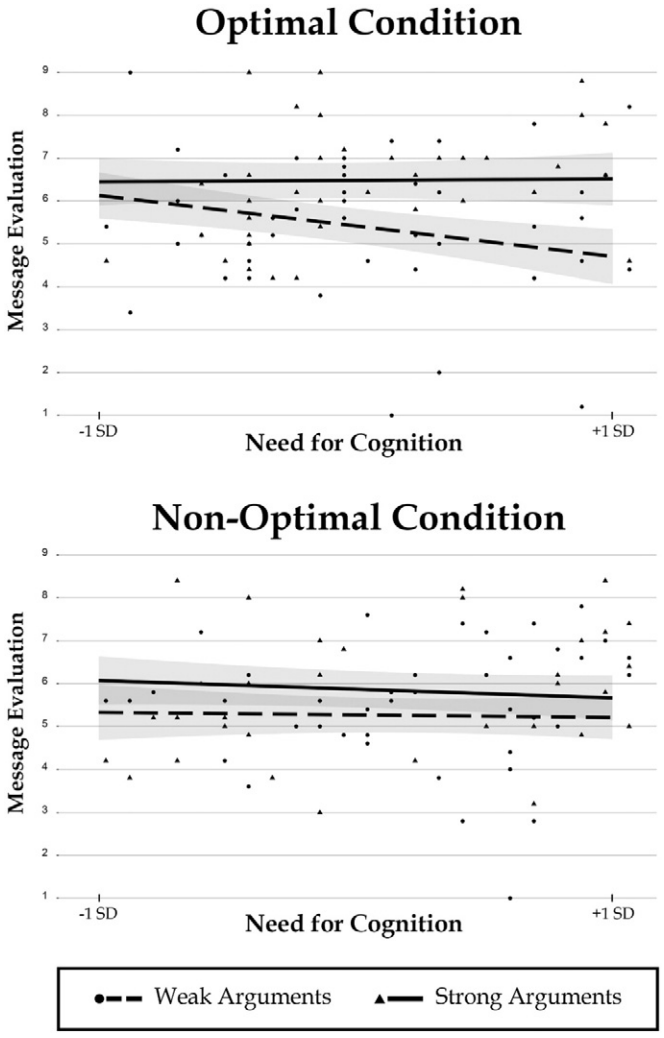


**Fig. 1.** A three-way interaction between need for cognition (NFC), argument quality (AQ), and the replication type on message evaluation. The key NFC × AQ interaction emerges only under the optimal conditions (a relatively long message and one that is relatively less personally relevant). Graphed lines are bound at one standard deviation above and below the mean of NFC. Data points from participants with more extreme NFC scores are not depicted. Confidence bands reflect 95% confidence intervals around estimates of message evaluation from the regression model where initial attitudes are set to the sample mean.

### 3.3. Using a reduced need for cognition scale

To test whether ML3's use of a reduced NFC scale may have impacted their ability to replicate CPM, data were submitted to the same model as before, replacing NFC with scores on the reduced (5 item) NFC scale.[11] Using this reduced scale also produces a significant three-way interaction, $B = 0.31$, $t(205) = 2.06$, $f^2 = 0.02$, $p = 0.04$, 95% CI: [0.01, 0.61], such that the NFC × AQ interaction remains nonsignificant in the non-optimal condition, $B = -0.15$, $f^2 = 0.002$, $p = 0.45$, but significant in the optimal condition, $B = 0.48$, $f^2 = 0.02$, $p = 0.04$, though the effect size is slightly smaller than when using the full scale.[12]

We also tested the NFC (5 item) × AQ interaction within the optimal conditions only, just as we did with the 18-item NFC scale. Results show that the two-way interaction is significant and somewhat larger than when tested in the context of the full model, $B = 0.48$, $t(103) = 2.13$, $f^2 = 0.04$, $p = 0.04$, 95% CI: [0.03, 0.93], but still smaller than when using the full scale ($f^2 = 0.07$). Overall, although the resulting effect sizes are smaller in analyses using the 5-item, rather than the 18-item, NFC scale, it does not appear that using a reduced scale was a critical factor in ML3's replication failure.

## 4. Discussion

This study aimed to compare a replication effort that was critiqued as non-optimal with conditions that were suggested as more optimal. We found that the NFC × AQ interaction first reported by CPM replicated under conditions that prior research and theory suggest are optimal for finding it. Under the non-optimal conditions used in the recent ML3 replication attempt, however, the NFC × AQ interaction did not emerge, effectively replicating the ML3 failure to replicate. Our strategy of showing that a replication effort can either succeed or fail depending on identified moderators (message length, personal relevance of topic) is relatively rare in the replication literature where only successes or failures are reported.

One aspect of ML3, however, did not seem to make an appreciable difference in the effect's replicability: the use of a reduced NFC scale. Although the 5-item NFC scale based on ML3's measure showed weaker internal reliability than the full 18-item scale and resulted in smaller effect sizes, it nonetheless significantly interacted with AQ under the optimal conditions. Put simply, using the larger NFC scale could not produce the NFC × AQ interaction under the non-optimal conditions, but using the short scale did not make the interaction go away under the optimal conditions. Nevertheless, it can be risky to use reduced scales in replication efforts. Indeed, other research has failed to produce predicted NFC effects using a 2-item version of the scale, but the effects emerged with a longer version (e.g., Kam, 2005; cf. Bakker & Lelkes, 2016).

A discussion of replication also warrants a comment on effect sizes and their consistency across different studies. Note that the effect sizes for the NFC × AQ interaction under the optimal conditions in the present study ($f^2 = 0.03$ from the full analysis; $f^2 = 0.07$ from the analysis of the optimal condition only) are larger than the effect size found by ML3 for the same interaction ($f^2 = 5.46\,e-5$), but smaller than the effect size originally found by CPM ($f^2 = 0.20$). Despite being smaller than the effect size from the original study, however, the NFC × AQ effect we found under optimal conditions is comparable to the meta-analytic effect size from 5 studies testing this hypothesis reported by Cacioppo et al. (1996; $f^2 = 0.07$) especially in the analyses of participants in the optimal condition only ($f^2 = 0.07$).[13] Furthermore, using the SPSS macros accompanying Smithson (2001), a 95% confidence interval could be computed around the effect size ($f^2$) in CPM, based on the *F*-value and degrees of freedom for the two-way interaction. The resulting confidence interval, [0.06, 0.41], includes the effect size found in the present study's analysis of the optimal condition only—the analysis most consistent with a straightforward replication.[14]

It is worth reiterating that the "optimal" condition in the present experiment was not intended to be identical to the materials and procedure of the original CPM experiment. We used longer messages, specified that the policy would take effect 10 years in the future, and used specific arguments that were different from CPM (whose arguments were no longer available). Also, the arguments were not specifically tailored to the participant population as they were in CPM. The current arguments were more comparable in content to ML3 and other persuasion research using the comprehensive exam topic. Any of these differences might explain why the effect size for the NFC × AQ interaction in our optimal condition is smaller than in the original experiment. In addition, the fact that we used an 18-item NFC scale rather than the original 34-item scale could also play a role as could the fact that we used NFC as a continuous measure rather than using the CPM procedure of relying on the upper and lower thirds of the distribution (see the Online supplement). However, our goal was not to replicate the original CPM study exactly, but was instead to instantiate optimal conditions for finding the interaction effect based not only on CPM but on developments in the literature since the original publication. The fact that our optimal conditions succeeded in producing the NFC × AQ interaction suggests that some of the procedures implemented by ML3 that did not follow CPM nor take advantage of the subsequent literature worked against finding the effect, though we cannot say which feature or features of our multifaceted optimal condition were responsible.

These results have clear implications for replication efforts more broadly. Given our evidence that original effects can replicate under optimal conditions suggested by existing theory and research, we advocate for the practice of *informed replication*—considering the full body of research conducted since the original study to inform replication efforts. This approach is generally lacking in contemporary replication programs. For example, commenting on the known moderators of priming effects, Dijksterhuis (2014) wrote: "it is unfortunate that these well-known moderators have not been taken into account in most replication efforts" (p. 74). Notably, however, moderators may not be explicit in the original study's report, and in fact, may not have been identified until *after* the original study was published. This means that even following an original procedure exactly or conceptually may overlook important features that have been documented

[11] Another way to address the previous discrepancies in the scales used to assess NFC is to run an analysis in which people in the non-optimal condition have an NFC score based on the reduced 5-item scale and people in the optimal condition have an NFC score based on the full 18-item scale. Overall, this had little effect on the conclusions that can be drawn from the data. The three-way interaction was still significant, $B = 0.38$, $t(205) = 2.44$, $p = 0.02$. The NFC (18 items) × AQ interaction in the optimal condition is significant, as in the previously reported analysis, $B = 0.62$, $t(205) = 2.49$, $p = 0.01$, and the NFC (5 items) × AQ interaction in the non-optimal condition is still nonsignificant, $B = -0.15$, $t(205) = -0.76$, $p = 0.44$.

[12] Another difference between ML3 and CPM is that the former treated NFC as a continuous variable and the latter treated it as categorical, focusing on individuals at the extreme ends of the scale. In the current data, there is a trend for larger effect sizes in the optimal condition when focusing on participants with extreme NFC scores (i.e., treating people who score in the lower and upper tertiles or quartiles of NFC as "low NFC" and "high NFC" groups, respectively). See Supplementary analyses for a full report.

[13] Although the present study focused on message evaluation as the dependent measure because it was the focus of ML3 and CPM, it is worth considering any possible differences between this approach and the other dominant approach in persuasion research: measuring people's evaluations of the *topic* of the message (rather than evaluations of the message itself). Notably, Cacioppo et al. (1996) also estimated the NFC × AQ effect using 11 studies that treated attitudes toward the topic of a message as the key dependent measure, and the resulting meta-analytic effect size is smaller ($f^2 = 0.02$) than when message evaluation is the dependent measure ($f^2 = .07$). This difference may be of interest to future attempts to demonstrate the NFC × AQ interaction.

[14] Cacioppo et al. (1996) did not report a confidence interval for their meta-analytic effect size estimate, nor did they report additional information to compute one. However, an analysis of the four other studies included in their meta-analysis shows lower bounds of the 95% confidence intervals around $f^2$ of 0.01, 0.00, 0.01, and 0.002, and upper bounds of 0.15, 0.09, 0.34, and 0.20, respectively. Thus, the effect size found in the present study is encompassed by all of these confidence intervals.

elsewhere. Thus, an informed replication is one that considers the full knowledge of an established literature and uses conditions under which an effect is known to occur. We encourage scholars disappointed by failures to replicate to conduct informed replications in which they compare optimal conditions with those implemented by the failed replication attempt.

## Acknowledgements

The authors would like to thank the members of the Attitudes and Persuasion Lab at Ohio State University and Thomas Vaughan-Johnston for helpful comments on this research. We also thank Victor Lee for his assistance with data collection.

## Appendix A. Supplementary materials

Supplementary materials, analyses, and raw data for this article can be found online at http://dx.doi.org/10.1016/j.jesp.2016.09.006.

## References

Bakker, B. N., & Lelkes, Y. (2016). *Selling ourselves short? How abbreviated measures of personality change the way we think about personality and politics.* (Unpublished manuscript), University of Amsterdam.

Brandt, M. J., Ijzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., & van't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. http://dx.doi.org/10.1016/j.jesp.2013.10.005.

Cacioppo, J. T., & Petty, R. E. (1979). Effects of message repetition and position on cognitive response, recall, and persuasion. *Journal of Personality and Social Psychology*, *37*, 97–109. http://dx.doi.org/10.1037/0022-3514.37.1.97.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116–131. http://dx.doi.org/10.1037/0022-3514.42.1.116.

Cacioppo, J. T., Petty, R. E., & Morris, K. (1983). Effects of need for cognition on message evaluation, argument recall, and persuasion. *Journal of Personality and Social Psychology*, *45*, 805–818. http://dx.doi.org/10.1037/0022-3514.45.4.805.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307. http://dx.doi.org/10.1207/s15327752jpa4803_13.

Cacioppo, J. T., Petty, R. E., Feinstein, J., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*, 197–253. http://dx.doi.org/10.1037/0033-2909.119.2.197.

Calanchini, J., Moons, W. G., & Mackie, D. M. (2016). Angry expressions induce extensive processing of persuasive appeals. *Journal of Experimental Social Psychology*, *64*, 88–98. http://dx.doi.org/10.1016/j.jesp.2016.02.004.

Carpenter, C. J. (2015). A meta-analysis of the ELM's argument quality × processing type predictions. *Human Communication Research*, *41*(4), 501–534. http://dx.doi.org/10.1111/hcre.12054.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, *9*(1), 40–48. http://dx.doi.org/10.1177/1745691613513470.

Clark, J. K., & Wegener, D. T. (2013). Message position, information processing, and persuasion: The Discrepancy Motives Model. In P. Devine, & A. Plant (Eds.), *Advances in experimental social psychology*, *Vol. 47*. (pp. 189–232). San Diego, CA: Academic Press. http://dx.doi.org/10.1016/B978-0-12-407236-7.00004-8.

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

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99. http://dx.doi.org/10.1016/j.jesp.2015.10.002.

Dijksterhuis, A. (2014). Welcome back theory! *Perspectives on Psychological Science*, *9*(1), 72–76. http://dx.doi.org/10.1177/1745691613513472.

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.

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68–80. http://dx.doi.org/10.1016/j.jesp.2015.07.009.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037a. http://dx.doi.org/10.1126/science.aad7243.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* New York: The Guilford Press.

Kam, C. D. (2005). Who toes the party line? Cues, values, and individual differences. *Political Behavior*, *27*(2), 163–182. http://dx.doi.org/10.1007/s11109-005-1764-y.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahnik, S., Bernstein, M. J., & Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*, 142–152. http://dx.doi.org/10.1027/1864-9335/a000017.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 943–951. http://dx.doi.org/10.1126/science.aac4716.

Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology*, *37*(10), 1915–1926. http://dx.doi.org/10.1037/0022-3514.37.10.1915.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change.* New York: Springer Verlag.

Petty, R. E., & Cacioppo J.T. (1990). Involvement and persuasion: Tradition versus integration. *Psychological Bulletin*, *107*, 367–374. http://dx.doi.org/10.1037/0033-2909.107.3.367.

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.

Petty, R. E., Harkins, S. G., & Williams, K. D. (1980). The effects of group diffusion of cognitive effort on attitudes: An information processing view. *Journal of Personality and Social Psychology*, *38*, 81–92. http://dx.doi.org/10.1037/0022-3514.38.1.81.

See, Y. H. M., Petty, R. E., & Evans, L. M. (2009). The impact of perceived message complexity and need for cognition on information processing and attitudes. *Journal of Research in Personality*, *43*, 880–889. http://dx.doi.org/10.1016/j.jrp.2009.04.006.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76–80. http://dx.doi.org/10.1177/1745691613514755.

Simonsohn, U. (2014). No-way interactions. *Data Colada*. http://dx.doi.org/10.15200/winn.142559.90552.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-Curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. http://dx.doi.org/10.1037/a0033242.

Smith, S. M., & Petty, R. E. (1996). Message framing and persuasion: A message processing analysis. *Personality and Social Psychology Bulletin*, *22*, 257–268. http://dx.doi.org/10.1177/0146167296223004.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*(4), 605–632. http://dx.doi.org/10.1177/00131640121971392.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*, 6454–6459. http://dx.doi.org/10.1073/pnas.1521897113.

Wheeler, S. C., Petty, R. E., & Bizer, G. Y. (2005). Self-schema matching and attitude change: Situational and dispositional determinants of message elaboration. *Journal of Consumer Research*, *31*, 787–797. http://dx.doi.org/10.1086/426613.

Widaman, K. F., Little, T. D., Preacher, K. J., & Sawalani, G. M. (2011). On creating and using short forms of scales in secondary research. In K. H. Trzesniewski, B. M. Donnellan, & R. E. Lucas (Eds.), *Secondary data analysis: An introduction for psychologists* (pp. 39–61). Washington, DC: American Psychological Association.