Check for updates

Special Section: Many Labs 5 Project
Empirical Article: Registered Report

Advances in Methods and Practices in Psychological Science
2020, Vol. 3(3) 394–404
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245918785165
www.psychologicalscience.org/AMPPS

Many Labs 5: Registered Multisite Replication of the Tempting-Fate Effects in Risen and Gilovich (2008)

Maya B. Mathur[1], Diane-Jo Bart-Plange[2], Balazs Aczel[3], Michael H. Bernstein[4,5], Antonia M. Ciunci[5], Charles R. Ebersole[2], Filipe Falcão[6], Kayla Ashbaugh[7], Rias A. Hilliard[8], Alan Jern[8], Danielle J. Kellier[9], Grecia Kessinger[10], Vanessa S. Kolb[5], Marton Kovacs[3], Caio Ambrosio Lage[11], Eleanor V. Langford[2], Samuel Lins[6], Dylan Manfredi[12], Venus Meyet[10], Don A. Moore[13], Gideon Nave[12], Christian Nunnally[14], Anna Palinkas[3], Kimberly P. Parks[2], Sebastiaan Pessers[15], Tiago Ramos[6], Kaylis Hase Rudy[10], Janos Salamon[3,16], Rachel L. Shubella[8], Rúben Silva[6], Sara Steegen[15], L. A. R. Stein[4,5,17], Barnabas Szaszi[3], Peter Szecsi[3], Francis Tuerlinckx[15], Wolf Vanpaemel[15], Maria Vlachou[15], Bradford J. Wiggins[10], David Zealley[10], Mark Zrubka[18], and Michael C. Frank[19]

[1]Quantitative Sciences Unit, Stanford University; [2]Department of Psychology, University of Virginia; [3]Institute of Psychology, ELTE Eötvös Loránd University; [4]Department of Behavioral and Social Sciences, School of Public Health, Brown University; [5]Department of Psychology, University of Rhode Island; [6]Department of Psychology, University of Porto; [7]Department of Biology and Biomedical Engineering, Rose-Hulman Institute of Technology; [8]Department of Humanities, Social Sciences, and the Arts, Rose-Hulman Institute of Technology; [9]Perelman School of Medicine, University of Pennsylvania; [10]Department of Psychology, Brigham Young University–Idaho; [11]Department of Psychology, Pontifical Catholic University of Rio de Janeiro; [12]Marketing Department, The Wharton School of the University of Pennsylvania; [13]Haas School of Business, University of California at Berkeley; [14]Department of Computer Science and Software Engineering, Rose-Hulman Institute of Technology; [15]Psychology and Educational Sciences, University of Leuven; [16]Doctoral School of Psychology, ELTE Eötvös Loránd University; [17]Rhode Island Training School, Rhode Island Department of Children, Youth and Families; [18]Department of Psychology, University of Amsterdam; and [19]Department of Psychology, Stanford University

Abstract

Risen and Gilovich (2008) found that subjects believed that "tempting fate" would be punished with ironic bad outcomes (a main effect), and that this effect was magnified when subjects were under cognitive load (an interaction). A previous replication study (Frank & Mathur, 2016) that used an online implementation of the protocol on Amazon Mechanical Turk failed to replicate both the main effect and the interaction. Before this replication was run, the authors of the original study expressed concern that the cognitive-load manipulation may be less effective when implemented online than when implemented in the lab and that subjects recruited online may also respond differently to the specific experimental scenario chosen for the replication. A later, large replication project, Many Labs 2 (Klein et al. 2018), replicated the main effect (though the effect size was smaller than in the original study), but the interaction was not assessed. Attempting to replicate the interaction while addressing the original authors' concerns regarding the protocol

Corresponding Author:
Maya B. Mathur, Quantitative Sciences Unit, 1701 Porter Dr., Palo Alto, CA 94305
E-mail: mmathur@stanford.edu

for the first replication study, we developed a new protocol in collaboration with the original authors. We used four university sites ($N = 754$) chosen for similarity to the site of the original study to conduct a high-powered, preregistered replication focused primarily on the interaction effect. Results from these sites did not support the interaction or the main effect and were comparable to results obtained at six additional universities that were less similar to the original site. Post hoc analyses did not provide strong evidence for statistical inconsistency between the original study's estimates and our estimates; that is, the original study's results would not have been extremely unlikely in the estimated distribution of population effects in our sites. We also collected data from a new Mechanical Turk sample under the first replication study's protocol, and results were not meaningfully different from those obtained with the new protocol at universities similar to the original site. Secondary analyses failed to support proposed substantive mechanisms for the failure to replicate.

Keywords
replication, reproducibility, heuristic, magical thinking, open data, open materials, preregistered

Received 12/17/17; Revision accepted 5/18/18

Risen and Gilovich (2008) examined the existence and mechanisms of the belief that "tempting fate" is punished with ironic bad outcomes. They hypothesized, for example, that students believe that they are more likely to be called on in class to answer a question about the assigned reading if they have not done the reading (and thus have tempted fate) than if they have come to class prepared (and thus have not tempted fate). Risen and Gilovich additionally hypothesized that deliberative thinking, also called *System 2 processing* (Epstein, Lipson, Holstein, & Huh, 1992), may help suppress irrational heuristics regarding tempting fate, and thus that a cognitive-load manipulation designed to preoccupy System 2 resources would magnify the effect of tempting fate on subjects' perception of the likelihood of a bad outcome. That is, Risen and Gilovich hypothesized a positive interaction effect of cognitive load and tempting fate on the perceived likelihood of an ironic bad outcome.

Risen and Gilovich's (2008) Study 6, the focus of the current replication study, used a between-subjects factorial design to test these hypotheses ($n = 120$ in the final sample). Subjects were randomly assigned to read a scenario in which they imagined themselves either having tempted fate by not having done the assigned reading or not having tempted fate by having done the assigned reading. Additionally, subjects' cognitive load was manipulated by random assignment. Subjects not under cognitive load simply read the scenario and then judged their likelihood of being called on in class (scale from 0 to 10). Subjects under cognitive load counted backward by 3s from a large number while reading the scenario and then provided the likelihood judgment. This study provided evidence for the predicted main effect of tempting fate in subjects not assigned to cognitive load. The estimated difference in perceived likelihood of being called on in class (b) was 1.03 on the 11-point scale, 95% confidence interval (CI) = [0.09, 1.97], $p = .03$[1]; the corresponding standardized mean difference (*SMD*) was 0.56, 95% CI = [0.05, 1.08]. This study also supported the *focal interaction* effect (i.e., the interaction between tempting fate and cognitive load); the difference in perceived likelihood was 1.54, 95% CI = [0.05, 3.03], $p = .04$, and the corresponding *SMD* was 0.75, 95% CI = [0.02, 1.47].

We selected Risen and Gilovich's (2008) Study 6 for our contribution to Many Labs 5 because, as required by the selection criteria for the Many Labs 5 replications, this study was subject to a previous replication attempt as part of the Open Science Collaboration's (2015) Reproducibility Project: Psychology (RP:P). The previous replication (Frank & Mathur, 2016) found little evidence for either the main effect of tempting fate without cognitive load ($N = 226$; $b = 0.20$, 95% CI = [–0.58, 0.97], $p = .62$; *SMD* = 0.09, 95% CI = [–0.26, 0.44]) or the focal interaction ($b = 0.03$, 95% CI = [–1.14, 1.20], $p = .96$; *SMD* = 0.01, 95% CI = [–0.51, 0.54]). However, prior to the data collection for this previous replication effort, the authors of the original study expressed concerns about the protocol. Specifically, the replication was implemented on the crowdsourcing website Amazon Mechanical Turk (MTurk), a setting that could potentially compromise the cognitive-load manipulation if subjects were already multitasking or were distracted. Additionally, the experimental scenario, which required subjects to imagine being unprepared to answer questions in class, may be less personally salient to subjects not enrolled in an elite university such as Cornell University, the site of the original study. Because of feasibility constraints and because the original report had not specifically described constraints on when the effect would occur, the RP:P replication proceeded without addressing these concerns. Thus, as part of the Many Labs 5 project, the present multisite replication (a)

Table 1. Comparison of the Experimental Protocols Used in the Original Study, the Reproducibility Project: Psychology (RP:P) Replication, and the Present Updated Replication

Original protocol (Risen & Gilovich, 2008, Study 6)	RP:P replication protocol (Frank & Mathur, 2016)	Updated replication protocol	Reason for update
Subjects were undergraduates at Cornell University.	Subjects were U.S. residents participating online through Amazon Mechanical Turk.	Subjects in the primary analyses were undergraduates at U.S. universities with median SAT scores above the 90th percentile nationally.	Subjects in settings with higher academic pressure may find the stimuli more personally salient. A university's average SAT score may serve as a proxy for academic pressure.
Subjects completed the experiment in a low-distraction, private lab setting.	No restrictions were placed on the physical setting in which subjects completed the experiment.	Subjects completed the experiment in controlled physical settings with reasonable isolation from other subjects (e.g., private lab room, private cubicles in a shared room).	The cognitive-load manipulation may be more effective when there are fewer other distractions.

reassessed the replicability of Risen and Gilovich's results using an updated protocol designed in collaboration with the original authors to mitigate potential problems with the RP:P replication protocol and (b) formally assessed the effect of updating the protocol in this manner by comparing its results with results from data newly collected under that previous protocol.

Concurrently with the present study, an independent group (Many Labs 2) conducted a multisite replication testing Risen and Gilovich's (2008) main effect (i.e., the effect of tempting fate in the absence of cognitive load), but not the interaction (Klein et al., 2018). Their primary analysis sample comprised undergraduates at universities and colleges in the United States and abroad ($N = 4{,}599$). These subjects judged their likelihood of being called on in class to be higher when they had tempted fate ($M = 4.61$, $SD = 2.42$) than when they had not tempted fate ($M = 4.07$, $SD = 2.36$), $t(4597) = 7.57$, $p = 4.4 \times 10^{-14}$; $SMD = 0.22$, 95% CI = [0.17, 0.28]. Thus, this replication provided strong evidence for a main effect of tempting fate, albeit an effect of smaller magnitude than in the original study. We discuss the results of the present study in light of these existing findings.

Disclosures

Preregistration

The protocol, sample-size criteria, exclusion criteria, and plan for statistical analysis were preregistered[2] at the Open Science Framework. The protocol can be accessed at https://osf.io/8y6st/, and the analysis plan can be accessed at https://osf.io/vqd5c/. Departures from these plans are reported in this article.

Data, materials, and online resources

All data, materials, and analysis code are publicly available at https://osf.io/h5a9y/. The Supplemental Material (http://journals.sagepub.com/doi/10.1177/2515245918785165) contains additional descriptive statistics for each site, methodological details of the statistical analyses, results of sensitivity analyses to verify the main results, and a formal comparison of our results with those of Many Labs 2 (Klein et al., 2018).

Reporting

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Ethical approval

Sites obtained ethics-committee approval when appropriate given their geographical location and institutional requirements, and data were collected in accordance with the Declaration of Helsinki.

Method

In consultation with the original authors, we designed a replication protocol that more closely duplicated the original design than did the RP:P replication (Table 1). Our primary analyses used only data from university sites that were located in the United States and that met a criterion for being similar to the original site with respect to academic competitiveness (i.e., median SAT score > 90th percentile nationally); we refer to these sites as *similar sites*. In addition, we used the previous

RP:P replication protocol without modification to collect data from a new sample on MTurk. Finally, we collected secondary data at several universities that did not meet the SAT criterion for similarity to Cornell or were located outside the United States; we refer to these sites as *dissimilar sites*. Data from the dissimilar sites were used in secondary analyses to increase power and assess whether, as hypothesized, similarity to the original test site in fact moderates the focal interaction. At sites where subjects were not expected to speak fluent English, questionnaire materials were translated and then verified through independent back-translation.

The primary effects we statistically estimated were (a) the focal interaction effect within the similar sites and (b) the difference in this interaction between the similar sites and MTurk (modeled as a three-way interaction, as described later). We also report the main-effect estimate to provide context for the focal interaction. Sample sizes were chosen to provide, in aggregate, more than 80% power to detect a three-way interaction with effect size (*SMD*) more than 0.75. Because detecting the three-way interaction required substantially larger sample sizes than detecting the focal interaction alone, this choice of sample sizes also provided more than 99% power to detect, within the similar sites alone, a focal interaction of the size reported in the original study. Researchers at each site attempted to reach these sample-size criteria within their own site, though in many cases this was not feasible. Site-level and aggregate analyses were conducted by the first author, who was blinded to results until all the sites had completed data collection; these analyses were audited for accuracy by other authors.

We collected four new measures, developed in discussion with the original authors, for use in secondary analyses. To check the effectiveness of the cognitive-load manipulation, we asked subjects assigned to cognitive load to use a scale from 0 to 10 to rate the perceived effort associated with the load task ("How much effort did the counting task require?") and the task's difficulty ("How difficult was the counting task?"). Additionally, the original authors speculated that the experimental scenario (regarding answering questions in class) may be more personally salient to subjects in an academically competitive environment similar to that of the original study than to MTurk subjects or subjects at dissimilar universities. To assess this possibility, we developed new measures that required subjects to evaluate the importance of answering questions correctly in class ("If you were a student in the scenario you just read about, how important would it be for you to answer questions correctly in class?") and how negative the experience of answering incorrectly would be to them ("If you were a student in the class, how bad would you feel if you were called on by the professor, but couldn't answer the question?"). These questions were also answered on a scale from 0 to 10.

Results

Descriptive analyses

Table 2 lists the sample sizes, the numbers of excluded subjects, and protocol characteristics for all the sites. To estimate the main effect of tempting fate and the focal interaction within each site, we fit an ordinary least squares regression model of the perceived likelihood of being called on in class; the predictors were tempting-fate condition, cognitive-load condition, and their interaction. This analysis approach is statistically equivalent to the analysis of variance model fit in the original study but also yields coefficient estimates that are directly comparable to those estimated in the primary analysis models, discussed in the next section. Figures 1 and 2, respectively, display these within-site estimates for the main effect and interaction.[3]

The main-effect estimates for three of the four similar sites were in the same direction as the original study's estimate, though of considerably smaller magnitude. The *p* values for the estimates at the similar sites ranged from .31 to .94. The main-effect estimate for the MTurk sample was in the same direction as the original study's estimate, but was of smaller size, and it was almost identical to the estimate previously obtained under the same protocol in RP:P (Frank & Mathur, 2016). Across all 10 university sites, 9 of the main-effect estimates were in the same direction as in the original study. However, these estimates were of smaller magnitude than the original except that the estimate at ELTE Eötvös Loránd University was comparable to that of the original study.

The focal-interaction estimates at two of the four similar sites were in the same direction as the original, but again were of considerably smaller magnitude. At the four similar sites, *p* values ranged from .43 to .89. For the MTurk sample, the interaction estimate was in the direction opposite that of the original estimate and was slightly larger in magnitude than the RP:P estimate. Overall, 4 of the 10 university sites had point estimates in the same direction as in the original study, but all of these estimates were of smaller magnitude. With one exception, *p* values across all the universities ranged from .21 to .99. At ELTE Eötvös Loránd University, the *p* value was .05 for a large point estimate in the direction opposite that of the original study's estimate.

Primary analyses

In the primary analyses, we (a) estimated the focal interaction under the updated protocol at the similar

Table 2. Summary of Sites and Participants

Site	Location	Subjects included in analyses (*n*)	Subjects excluded from analyses (*n*)	Recruitment and compensation	Language	Physical setting
Online site						
Amazon Mechanical Turk	—	2,973	162	Paid U.S. online workers	English	Online
Similar university sites						
University of Pennsylvania	Philadelphia, PA	335	24	Paid undergraduates from a university subject pool	English	Lab with private cubicles (groups of about 20)
University of California at Berkeley	Berkeley, CA	200	23	Undergraduate business majors given credit	English	Lab with private cubicles (groups of 1–13)
University of Virginia	Charlottesville, VA	151	5	Undergraduate introductory-psychology students given credit	English	Lab with private rooms (groups of 1–4)
Stanford University	Stanford, CA	68	1	Undergraduate introductory-psychology students given credit	English	Lab room (individually)
Dissimilar university sites						
ELTE Eötvös Loránd University	Budapest, Hungary	284	7	Undergraduates psychology students given credit	Hungarian	Lab with private cubicles (groups of 5–20)
University of Leuven	Leuven, Belgium	118	9	Undergraduates from a university subject pool given credit or pay	Dutch	Lab with private cubicles (groups of 1–2)
University of Porto	Porto, Portugal	91	13	Undergraduate introductory-psychology students given no compensation	Portuguese	Lab with private cubicles (groups of 1–4)
Brigham Young University–Idaho	Rexburg, ID	84	6	Undergraduate introductory-psychology students given credit and a raffle entry	English	Lab with private rooms (groups of 1–2)
University of Rhode Island	Kingston, RI	81	9	Undergraduate psychology students given credit	English	Lab with private cubicles (groups of 1–4)
Rose-Hulman Institute of Technology	Terre Haute, IN	56	2	Uncompensated peers of undergraduate research assistants	English	Lab room (individually)

Note: Subjects were excluded from analyses in keeping with a priori criteria; additional post hoc exclusions were followed at ELTE Eötvös Loránd University.

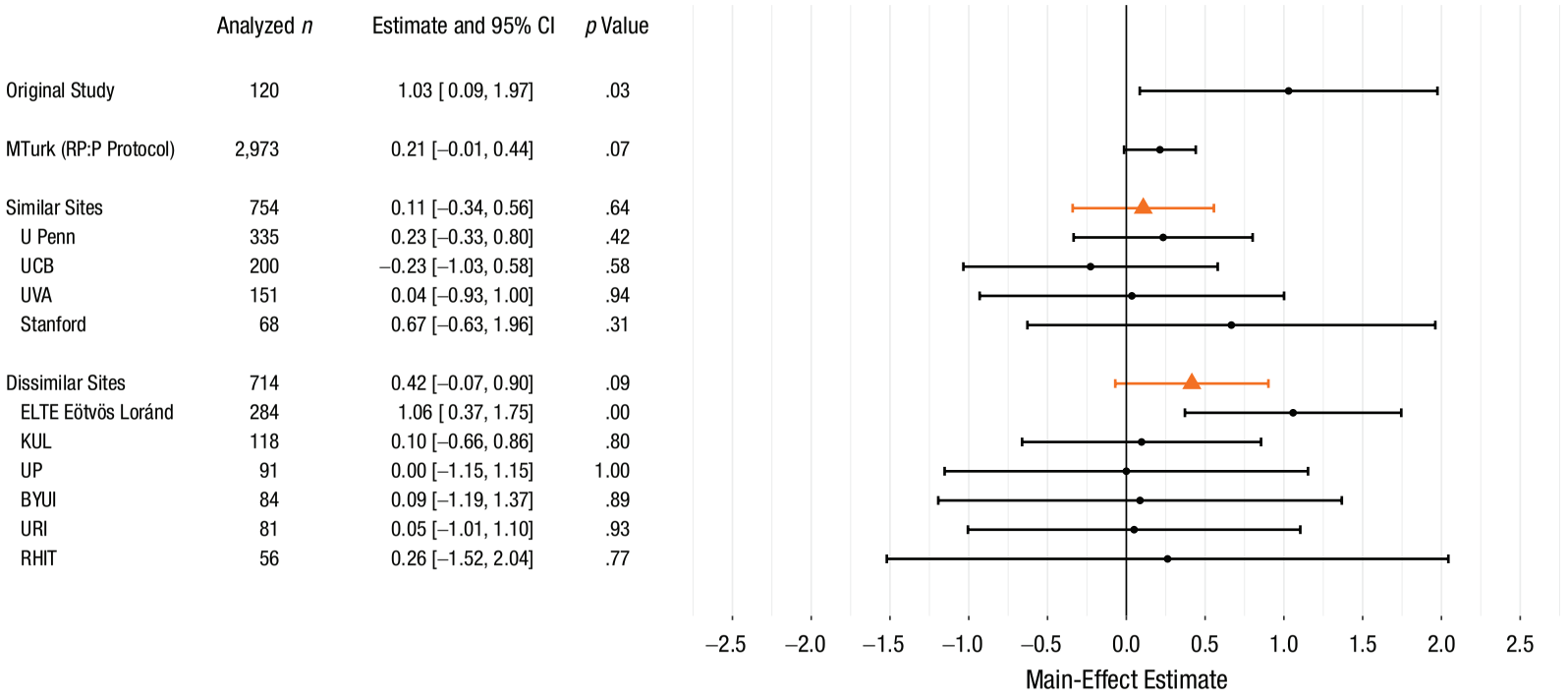

Fig. 1. Forest plot for the main-effect estimates from Risen and Gilovich's (2008) original study and from our replication project (ordered by site type and by sample size within site type). Point estimates and 95% confidence intervals (CIs) for the individual sites (black circles and bars) are from ordinary least squares regressions fit to each site's data. Pooled point estimates and 95% CIs (orange triangles and bars) are from the primary mixed model for the similar sites and from the secondary mixed model for the dissimilar sites. These pooled point estimates represent the average main effect among subjects in similar or in dissimilar universities. MTurk = Amazon Mechanical Turk; RP:P = Reproducibility Project: Psychology; U Penn = University of Pennsylvania; UCB = University of California at Berkeley; UVA = University of Virginia; KUL = University of Leuven; UP = University of Porto; BYUI = Brigham Young University–Idaho; URI = University of Rhode Island; RHIT = Rose-Hulman Institute of Technology.

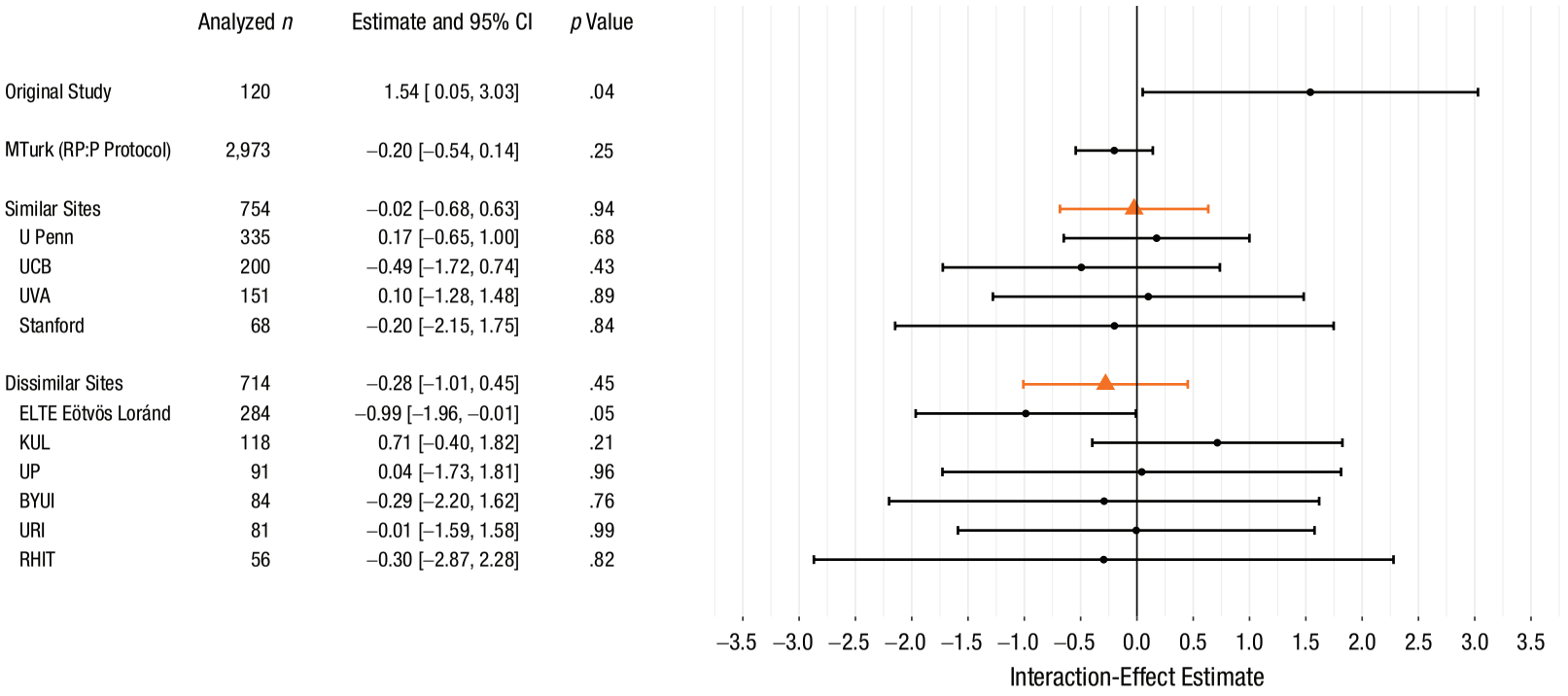

Fig. 2. Forest plot for the focal-interaction estimates from Risen and Gilovich's (2008) original study and from our replication project (ordered by site type and by sample size within site type). Point estimates and 95% confidence intervals (CIs) for the individual sites (black circles and bars) are from ordinary least squares regression fit to each site's data. Pooled point estimates and 95% CIs (orange triangles and bars) are from the primary mixed model for the similar sites and from the secondary mixed model for the dissimilar sites. These pooled point estimates represent the average interaction effect among subjects in similar or in dissimilar universities. MTurk = Amazon Mechanical Turk; RP:P = Reproducibility Project: Psychology; U Penn = University of Pennsylvania; UCB = University of California at Berkeley; UVA = University of Virginia; KUL = University of Leuven; UP = University of Porto; BYUI = Brigham Young University–Idaho; URI = University of Rhode Island; RHIT = Rose-Hulman Institute of Technology.

Table 3. Comparison of the Estimated Main Effect and Focal Interaction Effect at the Similar University Sites and Under the RP:P Protocol

Parameter	Estimate	95% CI	*p* value
Main effect of tempting fate in the MTurk sample	0.21	[–0.01, 0.43]	.06
Main effect of tempting fate at the similar sites	0.11	[–0.34, 0.56]	.64
Effect of site type (similar site vs. MTurk) on the main effect of tempting fate	–0.11	[–0.61, 0.39]	.68
Tempting Fate × Cognitive Load interaction in the MTurk sample	–0.20	[–0.53, 0.13]	.24
Tempting Fate × Cognitive Load interaction at the similar sites	–0.02	[–0.68, 0.63]	.94
Effect of site type (similar site vs. MTurk) on the Tempting Fate × Cognitive Load interaction	0.18	[–0.56, 0.91]	.64

Note: The estimates are in units of perceived likelihood of a bad outcome, on an 11-point scale. $N = 3{,}727$. The estimates and inference for Amazon Mechanical Turk (MTurk) were obtained from the main analysis model, so they differ negligibly from the results of the subgroup analyses presented in Figures 1 and 2. The Reproducibility Project: Psychology (RP:P) replication study was conducted on MTurk. CI = confidence interval.

sites and (b) assessed whether these estimates differed from those obtained using the RP:P protocol. To this end, we combined data from the similar sites and MTurk to fit a linear mixed model with fixed effects representing main effects and interactions of tempting fate, cognitive load, and protocol (the updated protocol at similar sites vs. the MTurk protocol). To account for correlation of observations within a site, the model also contained by-site random intercepts and random slopes for tempting fate, cognitive load, and their interaction; in all analyses, all random effects were assumed to be independently and identically normally distributed.[4] This model allowed us to estimate the effects within the similar sites and within MTurk and permitted formal assessment of the extent to which these effects differed (via the three-way interaction of protocol, tempting fate, and cognitive load). Details of the model specification and the interpretation of each coefficient of interest are provided in the preregistered protocol.

The model for the primary analysis included 3,727 subjects from the similar sites and MTurk. Results are summarized in Table 3. Consistent with the RP:P results, the present results for the MTurk sample did not strongly support the main effect of tempting fate, though the small point estimate agreed in direction with that of the original study. Results obtained under the updated protocol at the similar sites also did not strongly support this main effect, and updating the protocol did not appear to change the main-effect estimate. Furthermore, neither results from the new MTurk sample nor results under the updated protocol supported the focal interaction. As was the case for the main effect, updating the protocol did not meaningfully affect the estimate for the interaction. Both the main effect of tempting fate and the focal interaction appeared to be homogeneous across sites (in each case, estimated *SD* of the random slopes = 0).

We conducted post hoc secondary analyses (see the Supplemental Material available online) to assess the extent to which the replication findings were statistically consistent with the original study's findings, that is, whether it is plausible that the original study was drawn from the same distribution of effects as were the effects in our replication project (Mathur & VanderWeele, 2020). These analyses indicated that if the original findings were in fact drawn from the same distribution as the results from the similar sites, the probability of the main-effect estimate in the original study being at least as extreme as was actually observed would be .12. For the focal interaction, the corresponding probability would be .07. Thus, despite the considerable difference between the two studies' main-effect and interaction estimates, these secondary analyses do not provide strong evidence for inconsistency between the original study and our replications. These results largely reflect the original study's small sample size and corresponding low precision (Mathur & VanderWeele, 2020).

Secondary analyses: all university sites

The planned secondary analyses addressed the same questions as the primary analyses, but additionally incorporated data from the dissimilar university sites ($N = 4{,}441$). Site type was treated as a categorical variable (MTurk, similar university site, or dissimilar university site).[5] Additionally, these analyses formally estimated the difference in results between the similar and dissimilar sites. The results of the secondary analyses are summarized in Table 4. They did not support the main effect or the focal interaction in any site type. The main-effect estimate and the interaction estimate were both comparable between the dissimilar sites and the similar sites. The main effect and interaction appeared to be more heterogeneous across sites than in the primary analyses (estimated *SD* of the main-effect random slopes = 0.19; estimated *SD* of the interaction random slopes = 0.37).

Table 4. Comparison of the Estimated Main Effect and Focal Interaction Effect at the Similar University Sites, at the Dissimilar University Sites, and Under the RP:P Protocol

Parameter	Estimate	95% CI	*p* value
Main effect of tempting fate in the MTurk sample	0.21	[–0.22, 0.65]	.34
Main effect of tempting fate at the similar sites	0.08	[–0.40, 0.57]	.73
Main effect of tempting fate at the dissimilar sites	0.42	[–0.07, 0.90]	.09
Effect of site similarity (similar vs. dissimilar) on the main effect of tempting fate	–0.33	[–1.02, 0.36]	.35
Tempting Fate × Cognitive Load interaction in the MTurk sample	–0.20	[–1.00, 0.60]	.62
Tempting Fate × Cognitive Load interaction at the similar sites	0.01	[–0.73, 0.76]	.97
Tempting Fate × Cognitive Load interaction at the dissimilar sites	–0.28	[–1.01, 0.45]	.45
Effect of site similarity (similar vs. dissimilar) on the Tempting Fate × Cognitive Load interaction	0.29	[–0.75, 1.34]	.58

Note: The estimates are in units of perceived likelihood of a bad outcome, on an 11-point scale. $N = 4{,}441$. The estimates and inference for Amazon Mechanical Turk (MTurk) were obtained from the main analysis model, so they differ negligibly from the results of the subgroup analyses presented in Figures 1 and 2. The Reproducibility Project: Psychology (RP:P) replication study was conducted on MTurk. CI = confidence interval.

Evaluating proposed explanations for replication failure

Anticipating that results might differ between the similar and dissimilar sites, we planned to conduct secondary analyses assessing proposed explanations for the previous replication failure in RP:P. However, given that the results did not appear to differ between the similar and dissimilar sites, we decided post hoc to modify our original plan and pursue simplified secondary analyses. First, we considered the suggestion that it may not be possible to implement the cognitive-load manipulation reliably in an online setting because of, for example, competing distractions in subjects' uncontrolled environments (Rand, 2012). We therefore assessed the extent to which the efficacy of the cognitive-load manipulation differed between MTurk subjects and all university subjects by fitting a mixed model with a three-way interaction of tempting fate, cognitive load, and an indicator for whether a subject completed the experiment on MTurk or at a university. The results suggested that the magnitude of the focal interaction between cognitive load and tempting fate was nearly identical for MTurk subjects and university subjects ($N = 4{,}441$, $b = -0.03$, 95% CI = [–0.99, 0.93], $p = .95$), although the fairly wide confidence interval indicated moderate uncertainty.

Second, to assess the effectiveness of the cognitive-load manipulation, we used data from subjects assigned to cognitive load and fit separate linear mixed models regressing perceived effort required for the counting task ($n = 1{,}852$) and perceived difficulty of the task ($n = 1{,}848$) on an indicator for whether a subject was recruited on MTurk or from a university.[6] If, as hypothesized, the cognitive-load manipulation was less effective on MTurk than in university settings, perceived effort or difficulty might have been lower for MTurk subjects than for university subjects. In fact, perceived effort associated with the cognitive-load task was comparable for MTurk and university subjects ($b = 0.63$, 95% CI = [–0.42, 1.68], $p = .24$), as was perceived difficulty ($b = 0.51$, 95% CI = [–0.11, 1.14], $p = .11$). Ultimately, these results do not suggest that the cognitive-load manipulation was less effective when implemented online rather than in person.

Third, to assess differences in academic attitudes as an explanation for the replication failure, we used data from subjects from all the sites[7], including MTurk, and fit linear mixed models regressing perceived importance of answering questions correctly ($n = 4{,}175$) and perceived negativity of answering questions incorrectly ($n = 4{,}172$) on site type (similar sites, dissimilar sites, or MTurk) with random intercepts by site. Contrary to expectation, MTurk subjects reported attaching somewhat greater importance to answering questions correctly than did subjects at the similar universities ($b = 1.02$, 95% CI = [0.45, 1.59], $p = 5 \times 10^{-4}$) or at the dissimilar universities ($b = 0.76$, 95% CI = [0.24, 1.29], $p = .005$). Additionally, MTurk subjects' responses when asked how bad it would be to answer incorrectly were comparable to the responses of subjects at the similar sites ($b = -0.03$, 95% CI = [–0.52, 0.45], $p = .89$) and at the dissimilar sites ($b = 0.45$, 95% CI = [0.01, 0.90], $p = .05$).

Finally, in a planned analysis, we assessed whether results varied with a site's similarity to Cornell, redefining similarity using a continuous proxy (namely, a university's estimated median total SAT score in 2018) rather than the dichotomous "similar" versus "dissimilar" criterion used for the primary analyses. Subjects from universities outside the United States and subjects from MTurk were excluded from this analysis, which left an analyzed sample of 975 students from seven universities with median SAT scores ranging from 1182 to 2178 of

2400 possible points. We assumed that universities with higher SAT scores would be more similar to Cornell (median SAT = 2134) and therefore considered a linear effect of median SAT score as a moderator of the main effect and the interaction between tempting fate and cognitive load. A mixed model did not suggest that median SAT score moderated either the main effect (b = 0.00 for a 10-point increase in SAT score, 95% CI = [−0.01, 0.02], p = .83) or the interaction (b = 0.00, 95% CI = [−0.02, 0.02], p = .97).

Comparison With the Results of Many Labs 2

As mentioned in the introduction, an independent multisite replication of Risen and Gilovich's (2008) experiment (Many Labs 2; Klein et al., 2018) found strong evidence for a main effect of tempting fate that was in the same direction as the effect obtained in the original study, but of smaller size; this finding stands in contrast to the present study's negligible main-effect estimate in the similar sites. We corresponded with the lead author of Many Labs 2 to identify protocol differences that might explain the discrepant results. We identified minor differences in questionnaire design, sampling frame, and statistical analysis (see Table S4 in the Supplemental Material) and pursued post hoc, exploratory analyses to gauge whether these differences were likely to explain the discrepant results (see the Supplemental Material for details). We compared results between the two studies' U.S. MTurk samples to identify effects of questionnaire design while holding constant the subject population. These analyses yielded nearly identical point estimates, suggesting that the differences in questionnaire design were not likely to have caused the discrepancy in the results for the full samples of the two replication projects. Additionally, we reanalyzed the raw data from Many Labs 2 in order to eliminate differences in statistical analyses, and we reanalyzed the raw data after redefining the studies' sampling frames to be more directly comparable. However, the two replication projects' point estimates remained meaningfully different in these analyses, so the source of the discrepancy remains unclear. Finally, we estimated a pooled main effect across the primary-analysis sites in both studies. This estimate provided strong evidence for a small main effect of tempting fate (see the Supplemental Material).

Conclusion

We used an updated protocol to replicate Risen and Gilovich (2008)'s Study 6 in controlled lab settings at universities chosen for their similarity to the original site. We additionally conducted replications on MTurk, as in the previous replication study (Frank & Mathur, 2016), as well as at universities less similar to the original site. This replication project has limitations: First, because the number of similar sites was small, we could not reliably assess variation in results across these sites. Second, as is true for all direct replications, our replication was limited to a single operationalization of the target effect; our results do not necessarily generalize to other experimental scenarios, for example.

Using the updated protocol at similar sites, we obtained a negligible estimate for the focal interaction between tempting fate and cognitive load. We also obtained a negligible estimated main effect of tempting fate in the absence of cognitive load. Results obtained under the updated protocol at the similar sites did not appear to differ from results obtained under the previous replication protocol on MTurk, nor did they differ meaningfully from the results obtained at the dissimilar universities. Secondary analyses did not support proposed explanations for the previous replication failure (namely, reduced effectiveness of the cognitive-load manipulation on MTurk or reduced personal salience of the experimental scenario on MTurk). Post hoc analyses of the main-effect and interaction estimates did not provide compelling evidence for statistical inconsistency between the original study and the replication under the original protocol. Ultimately, our results fail to support the tempting-fate effect and the interaction with cognitive load and also fail to support proposed substantive mechanisms for the replication failure in RP:P (Frank & Mathur, 2016). However, it is important to note that our negligible main-effect estimate stands in contrast to the small main effect identified in Many Labs 2 (Klein et al., 2018), and extensive post hoc analyses were unable to identify the source of the discrepancy.

Transparency

Action Editor: Daniel J. Simons
Editor: Daniel J. Simons
Author Contributions

C. R. Ebersole conceived the overall Many Labs 5 project. M. B. Mathur, C. R. Ebersole, and M. C. Frank designed this multisite replication study. M. B. Mathur and D.-J. Bart-Plange oversaw its administration. M. B. Mathur planned and conducted the statistical analyses (with code audited by M. C. Frank) and wrote the manuscript. The remaining authors collected data and audited site-level analyses. All the authors approved the final submitted manuscript, with one exception (sadly, S. Pessers passed away before the draft was written).

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported by a grant from the Association for Psychological Science and from Arnold Ventures. M. H. Bernstein was supported by Grant F31AA024358 from the National Institute on Alcohol Abuse and Alcoholism. G. Nave was supported by the Wharton Behavioral Lab.

Open Practices

Open Data: https://osf.io/h5a9y/
Open Materials: https://osf.io/h5a9y/
Preregistration: https://osf.io/8y6st/, https://osf.io/vqd5c/
All data, analysis scripts, and materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/h5a9y/. The protocol and analysis plans were preregistered at the Open Science Framework and can be accessed at https://osf.io/8y6st/ and https://osf.io/vqd5c/, respectively. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245918785165. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.



ORCID iD

Maya B. Mathur https://orcid.org/0000-0001-6698-2607

Acknowledgments

We are grateful to Jane Risen and Thomas Gilovich for their thoughtful and responsive feedback through the process of designing the protocol. We thank Sara Rose Christodoulou and Kate Kelly for scheduling assistance, Kimberly Marion and Jessica Simms for data-collection assistance, and Robert Botto for technical assistance. We thank Richard Klein for helpful discussions regarding the results of Many Labs 2 and for providing early access to the manuscript and raw data for that project.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/10.1177/2515245918785165

Notes

1. Approximate effect sizes were recomputed from rounded values reported by Risen and Gilovich (2008).
2. One site (Brigham Young University–Idaho) was permitted to collect data prior to preregistration of the statistical analysis plan because of their time constraints; the lead investigator and all other authors remained blinded to this site's results until preregistration and data collection were complete.
3. An alternative to using study-specific estimates would be to use estimates of the appropriate random slopes by site from the mixed model, but we report subset analyses for a descriptive characterization that relaxes the across-site distributional assumptions of the mixed model.
4. In a planned sensitivity analysis, we also refitted the same analysis of variance model used in the original study, which ignores correlation of observations within sites. This analysis yielded qualitatively similar results (see the Supplemental Material). We also obtained very similar results in additional sensitivity analyses in which we fitted a model to only the subset of data from the similar sites (dropping the MTurk coefficient) and in which we fitted meta-analytic counterparts to the primary model (see the Supplemental Material).
5. An alternative model specification in which all the universities were treated as a single category yielded similar results (see the Supplemental Material).
6. Because of an error in data collection, the new measures for perceived effort and difficulty were omitted at one site (University of California at Berkeley); thus, the subjects at this site were excluded in these analyses.
7. These analyses again excluded subjects at the University of California at Berkeley, because the new measures were not administered there.

References

Epstein, S., Lipson, A., Holstein, C., & Huh, E. (1992). Irrational reactions to negative outcomes: Evidence for two conceptual systems. *Journal of Personality and Social Psychology*, *62*, 328–339.

Frank, M. C., & Mathur, M. (2016). *Replication of Risen & Gilovich (2008, JPSP)*. Retrieved from https://osf.io/sg3su/

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490. doi:10.1177/2515245918810225

Mathur, M. B., & VanderWeele, T. J. (2020). New statistical metrics for multisite replication projects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. Advance online publication. doi:10.1111/rssa.12572

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

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172–179.

Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology*, *95*, 293–307.