Many Labs 5: Registered multisite replication of tempting-fate effects in Risen & Gilovich (2008)

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

[1] Harvard University, Boston, MA, United States

[2] Stanford University, Stanford, CA, United States

[3] University of Virginia, Charlottesville, VA, United States

[4] ELTE Eötvös Loránd University, Budapest, Hungary

[5] University of Rhode Island, Kingston, RI, United States

[6] University of Porto, Porto, Portugal

[7] Rose-Hulman Institute of Technology, Terre Haute, IN, United States

[8] Brigham Young University - Idaho, Rexburg, ID, United States

[9] Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil

[10] University of Pennsylvania, Philadelphia, PA, United States

[11] University of California at Berkeley, Berkeley, CA, United States

[12] University of Leuven, Belgium

[13] Brown University, Providence, RI, United States

[14] Rhode Island Training School, Cranston, RI, United States

Author Note

Correspondence concerning this article should be addressed to Maya B. Mathur, Quantitative Sciences Unit, 1070 Arastradero Rd, Palo Alto, CA, 94305. E-mail: mmathur@stanford.edu

Abstract

Risen & Gilovich (2008) found that subjects believe that "tempting fate" will be punished with ironic bad outcomes (a main effect) and that this effect is magnified under cognitive load (an interaction). A previous replication project (Open Science Collaboration, 2015) failed to replicate both the main effect and the interaction in an online implementation of the protocol that used Amazon Mechanical Turk. 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 and that subjects recruited online may respond differently to the specific experimental scenario chosen for replication. A later, large replication project (Many Labs 2) replicated the main effect (though the effect size was smaller than in the original study), but did not test for an interaction. To attempt to replicate the interaction while addressing the original authors' concerns regarding the 2015 protocol, we developed a new protocol in collaboration with the original authors. We used 4 university sites ($n = 754$ total) chosen for similarity to the site of the original study to conduct a high-powered, preregistered replication focused primarily on the interaction effect. Results did not support the focus interaction or the main effect and were comparable in 6 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 the replications; that is, the original study's results would not have been extremely unlikely in the estimated distribution of the replications. We also collected a new Mechanical Turk sample under the previous replication protocol, indicating that the updated protocol (i.e., conducting the study in person and in universities similar to the original site) did not meaningfully change replication results. Secondary analyses failed to support substantive mechanisms for the failure to replicate.

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

Many Labs 5: Registered multisite replication of tempting-fate effects in Risen & Gilovich (2008)

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, in fact, they had not done the reading (and thus had "tempted fate") versus if they had come to class prepared (and thus had not "tempted fate"). Risen and Gilovich (2008) additionally hypothesized that deliberative thinking (sometimes termed "System 2" processing (Epstein, Lipson, Holstein, & Huh, 1992)) may help suppress irrational heuristics regarding tempting fate, and thus a cognitive load manipulation designed to preoccupy System 2 resources would magnify the effect of tempting fate on subjects' perceived likelihood of a bad outcome. That is, they hypothesized a positive interaction between cognitive load and tempting fate on subjects' perceived likelihood of an ironic bad outcome.

Risen and Gilovich (2008)'s Study 6, the focus of replication, used a between-subjects factorial design to assess this possibility (total analyzed $n = 120$). Subjects were randomly assigned to read a scenario in which they imagined themselves having tempted fate by not having done the assigned reading or, alternatively, not having tempted fate by having done the assigned reading. Additionally, subjects were randomly assigned to complete the task with or without cognitive load. Subjects not under cognitive load simply read the scenario and then judged the likelihood of being called on in class. Subjects under cognitive load counted backwards by 3s from a large number while reading the scenario, after which they provided the likelihood judgment. This study provided evidence for the predicted main effect of tempting fate in subjects not assigned to cognitive load (estimated difference in perceived likelihood on a 0-10 scale after tempting fate vs. not tempting fate: $b = 1.03$ with 95% CI: [0.09, 1.97]; $p = 0.03$)[1] as well as the focus interaction effect (estimated effect of tempting

[1] Approximate effect sizes were recomputed from rounded values in Risen and Gilovich (2008).

fate vs. not tempting fate for subjects under cognitive load vs. not under cognitive load: $b =$ 1.54 with 95% CI: [0.05, 3.03]; $p = 0.04$).

We selected Risen and Gilovich (2008) for replication because, per the selection criteria of all Many Labs 5 replications, this study was subject to a previous replication attempt by Open Science Collaboration (2015). The previous replication found little evidence for either a main effect of tempting fate without cognitive load ($n = 226$, $b = 0.20$ with 95% CI: [-0.58, 0.97]; $p = 0.62$) or the focus interaction ($b = 0.03$ with 95% CI: [-1.14, 1.20]; $p = 0.96$) (Mathur & Frank, 2012). However, prior to the collection of replication data by this previous replication effort (termed "RPP"), the authors of the original study expressed concerns about the replication protocol. Due to feasibility constraints, the RPP replication proceeded without addressing these concerns. Specifically, the replication was implemented on the crowdsourcing website Amazon Mechanical Turk, 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 similar to Cornell University, the site of the original study. Thus, as part of the Many Labs 5 project, the present multisite replication aimed to: (1) reassess replicability of Risen and Gilovich (2008) using an updated protocol designed in collaboration with the original authors to mitigate potential problems with the previous replication protocol; and (2) formally assess the effect of updating the protocol in this manner by comparing its results to newly collected results under the previous replication protocol.

Concurrently with the present study, an independent group (Many Labs 2) conducted a multisite replication of the main effect, but not the interaction (Klein, 2017). Their primary analysis sample comprised undergraduates at universities and colleges in the United States and abroad ($n = 4599$). These subjects judged the likelihood of being called on to be higher when they had tempted fate (mean = 4.61, SD =2.42) than when they had not tempted fate (mean = 4.07, SD = 2.36; $t(4597) = 7.57$, $p = 4.4 \times 10^{-14}$, $d = 0.22$, 95% CI

$[0.17, 0.28]$), providing strong evidence for a main effect of tempting fate, albeit of smaller magnitude than in the original study. We discuss the results of the present study in light of these existing findings.

Disclosures

The protocol, sample size criteria, exclusion criteria, and statistical analysis plan were preregistered[2] with details publicly available (https://osf.io/8y6st/ for the protocol and https://osf.io/vqd5c/ for the analyses); departures from these plans are reported in this manuscript. All data, materials, and analysis code are publicly available and documented (https://osf.io/h5a9y/). Sites obtained ethics committee approval when appropriate to their geographical location and institutional requirements, and data were collected in accordance with the Declaration of Helsinki.

Methods

In consultation with the original authors, we designed a replication procotol that more closely duplicated the original design than did the RPP replication (Table 1). Primary analyses used only data from university sites located in the United States and meeting an academic criterion for similarity to the original site (Table 1, row 1); these sites are termed "similar sites". We additionally used the previous RPP replication protocol without modification to collect a new sample on Amazon Mechanical Turk ("MTurk"). Finally, we collected secondary data in several universities not meeting the SAT criterion for similarity to Cornell or located outside the United States, henceforth termed "dissimilar sites". Data from dissimilar sites were used in secondary analyses to further increase power and assess whether, as hypothesized, site similarity in fact moderates the focus effect. For sites whose subjects were not expected to speak fluent English, questionnaire materials were translated

[2] One site (BYUI) was permitted to collect data prior to preregistration of the statistical analysis plan due to their time constraints; the lead investigator and all other authors remained blinded to this site's results until preregistration and data collection were complete.

and verified through independent back-translation.

Original protocol	RPP replication protocol	Updated replication protocol	Reason for update
Subjects were undergraduates at Cornell University.	Subjects were United States residents participating online through Amazon Mechanical Turk.	Subjects in primary analyses were undergraduates at United States universities with median SAT scores >90th percentile nationally.	Subjects in settings with high academic pressure may find the stimuli more personally salient. A university's average SAT score may serve as a proxy for such 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 in all analyses 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 other distractions are minimal.

Table 1: Comparison of experimental protocols used in the original study, the RPP replication, and the present replication.

The primary statistical estimands were (1) the focus interaction within similar sites and (2) the difference in this interaction between similar sites and MTurk (modeled as a three-way interaction, as described below). Sample sizes were chosen to provide, in aggregate, more than 80% power to detect a three-way interaction with effect size more than 0.75 standard deviations of perceived likelihood. Because detecting the three-way interaction requires substantially larger sample sizes than detecting the focus interaction alone, this choice of sample sizes also provided $> 99\%$ power to detect, within similar sites alone, a focus interaction of the size reported in the original study. Each site additionally attempted to reach these power criteria internally, though in many cases this was not feasible. Site-level and aggregate analyses were conducted by one author (MBM), who was blinded to results until all 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. As manipulation checks for the effectiveness of the cognitive load manipulation, we asked subjects assigned to cognitive load to assess on a 0-10 scale the perceived effort associated with this 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 personally salient to subjects in an academically competitive environment similar to the site of the original study, but may be less so for MTurk subjects or subjects in dissimilar universities. To assess this possibility, we developed new measures in collaboration with the original authors which required subjects to evaluate on a 0-10 scale 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 the perceived negativity of answering incorrectly (*"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?"*).

Results

Descriptive analyses

Table 2 displays sample sizes, the number of exclusions, and protocol characteristics for all sites. To estimate the main effect of tempting fate and the focus interaction within each site, we fit an ordinary least squares regression model of perceived likelihood on tempting fate, cognitive load, and their interaction within each site. This analysis approach is statistically equivalent to the ANOVA model fit in the original study while also yielding coefficient estimates that are directly comparable to those estimated in primary analysis models, discussed below. Figures 1 and 2, respectively, display these within-site estimates for the main effect and interaction.[3]

Among the 4 similar sites, 3 had main effect estimates in the same direction as the original study estimate, albeit of considerably smaller magnitude. Main effect estimates in similar sites had p-values ranging from 0.31 to 0.94. In the MTurk sample, the main effect estimate was in the same direction as the original, but was of smaller size, and it was almost

[3] An alternative for the study-specific estimates would be to use estimates of random intercepts and random slopes by site from the mixed model, but here we use subset analyses for a descriptive characterization that relaxes the across-site distributional assumptions of the mixed model.

Site	Location	Analyzed *n*	Excluded *n*	Recruitment and compensation	Language	Physical setting
Online site						
Amazon Mechanical Turk (MTurk)	N/A	2973	162	U.S. online workers (pay)	English	Online
Similar university sites						
University of Pennsylvania (UPenn)	Philadelphia, PA	335	24	Undergraduates from university subject pool (pay)	English	Lab with private cubicles (groups of about 20)
University of California at Berkeley (UCB)	Berkeley, CA	200	23	Undergraduate business majors (credit)	English	Lab with private cubicles (groups of 1-13)
University of Virginia (UVA)	Charlottesville, VA	151	5	Undergraduates from introductory psychology class (credit)	English	Lab with private rooms (groups of 1-4)
Stanford University	Stanford, CA	68	1	Undergraduates from introductory psychology class (credit)	English	Lab room (individually)
Dissimilar university sites						
Eotvos Lorand University	Budapest, Hungary	284	7	Undergraduates from psychology course (credit)	Hungarian	Lab with private cubicles (groups of 5-20)
Katholieke Universiteit Leuven (KUL)	Leuven, Belgium	118	9	Undergraduates from university subject pool (credit or pay)	Dutch	Lab with private cubicles (groups of 1-2)
University of Porto (UP)	Porto, Portugal	91	13	Undergraduates from introductory psychology class (no compensation)	Portuguese	Lab with private cubicles (groups of 1-4)
Brigham Young University - Idaho (BYUI)	Rexburg, ID	84	6	Undergraduates from introductory psychology class (credit and raffle entry)	English	Lab with private rooms (groups of 1-2)
University of Rhode Island (URI)	Kingston, RI	81	9	Undergraduates from multiple psychology courses	English	Lab with private cubicles (groups of 1-4)
Rose-Hulman Institute of Technology (RHIT)	Terre Haute, IN	56	2	Recruited peers of undergraduate research assistants (no compensation)	English	Lab room (individually)

Analyzed n = total subjects included in analysis; excluded n = total subjects excluded from analysis in keeping with a priori criteria or post hoc exclusions at Eotvos Lorand University.

Table 2: Summary of sites and participants.

identical to the estimate previously obtained under the same protocol in RPP. Considering all 10 university sites, 9 had main effect estimates in the same direction as the original study. However, these estimates were of smaller magnitude than the original estimate with the exception of Eotvos Lorand University, which estimated a main effect comparable to that of the original study.

Considering the focus interaction estimate, 2 of 4 similar sites had estimates in the same direction as the original, and again, these were of considerably smaller magnitude with p-values ranging from 0.43 to 0.89. In the MTurk sample, the interaction estimate was in the

	Analyzed n	Estimate and 95% CI	
Original study	120	1.03 [0.09, 1.97]	0.03
RPP protocol			
MTurk	2973	0.21 [-0.01, 0.44]	0.07
Similar sites	754	0.11 [-0.34, 0.56]	0.64
U Penn	335	0.23 [-0.33, 0.80]	0.42
UCB	200	-0.23 [-1.03, 0.58]	0.58
UVA	151	0.04 [-0.93, 1.00]	0.94
Stanford	68	0.67 [-0.63, 1.96]	0.31
Dissimilar sites	714	0.42 [-0.07, 0.90]	0.09
Eotvos	284	1.06 [0.37, 1.75]	0.00
KUL	118	0.10 [-0.66, 0.86]	0.80
UP	91	0.00 [-1.15, 1.15]	1.00
BYUI	84	0.09 [-1.19, 1.37]	0.89
URI	81	0.05 [-1.01, 1.10]	0.93
RHIT	56	0.26 [-1.52, 2.04]	0.77

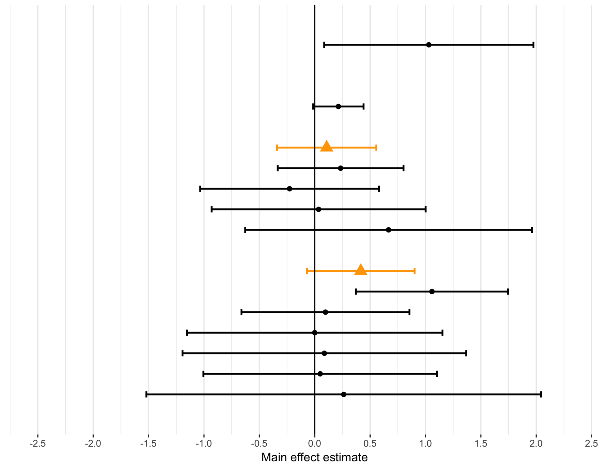


Figure 1: Forest plot for main effect estimates ordered by site type (MTurk, similar, dissimilar) and then by sample size. Point estimates and 95% CIs for each site (black circles) are from ordinary least squares regression fit to that site's data. For similar sites, pooled point estimates and 95% CIs (orange diamonds) are from the primary mixed model. For dissimilar sites (orange diamonds), these are from the secondary mixed model. Pooled point estimates represent the average main effect among subjects in similar universities or in all universities.

opposite direction from the original estimate and was slightly larger in magnitude than the RPP estimate. Considering all 10 university sites, 4 had point estimates in the same direction as the original study, all of which were of smaller magnitude. With one exception (Eotvos Lorand University with $p = 0.05$), p-values across all universities ranged from 0.21 to 0.99. Eotvos Lorand University obtained a large point estimate in the opposite direction from the original study.

Primary analyses

Primary analyses aimed to: (1) estimate the focus interaction and the main effect under the updated protocol in similar sites; and (2) assess whether the focus interaction and the main effect estimates differed between the updated protocol and the RPP 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 of tempting fate, cognitive load, and protocol

	Analyzed n	Estimate and 95% CI	p-value
Original study	120	1.54 [0.05, 3.03]	0.04
RPP protocol			
MTurk	2973	-0.20 [-0.54, 0.14]	0.25
Similar sites	754	-0.02 [-0.68, 0.63]	0.94
U Penn	335	0.17 [-0.65, 1.00]	0.68
UCB	200	-0.49 [-1.72, 0.74]	0.43
UVA	151	0.10 [-1.28, 1.48]	0.89
Stanford	68	-0.20 [-2.15, 1.75]	0.84
Dissimilar sites	714	-0.28 [-1.01, 0.45]	0.45
Eotvos	284	-0.99 [-1.96, -0.01]	0.05
KUL	118	0.71 [-0.40, 1.82]	0.21
UP	91	0.04 [-1.73, 1.81]	0.96
BYUI	84	-0.29 [-2.20, 1.62]	0.76
URI	81	-0.01 [-1.59, 1.58]	0.99
RHIT	56	-0.30 [-2.87, 2.28]	0.82

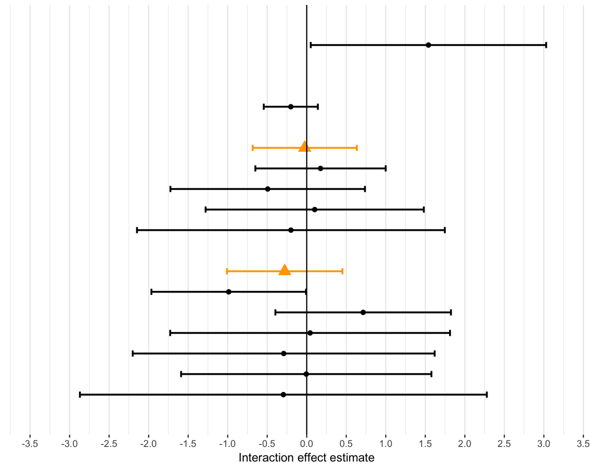


Figure 2: Forest plot for interaction estimates ordered by site type (MTurk, similar, dissimilar) and then by sample size. Point estimates and 95% CIs for each site (black circles) are from ordinary least squares regression fit to that site's data. For similar sites, pooled point estimates and 95% CIs (orange diamonds) are from the primary mixed model. For dissimilar sites (orange diamonds), these are from the secondary mixed model. Pooled point estimates represent the average interaction effect among subjects in similar universities or in all universities.

(similar sites under the updated protocol vs. MTurk). To account for correlation of observations within a site, the model also contained random intercepts by site and random slopes by site of tempting fate, cognitive load, and their interaction; in all analyses, all random effects were assumed independently and identically normal.[4] This model allows estimation of the focus effect within similar sites and within MTurk and permits formal assessment of the extent to which these effects differ (via the three-way interaction of protocol, tempting fate, and cognitive load). Details of the model specification and

[4]As a planned sensitivity analysis, we also refit the same ANOVA model used in the original study, which ignores correlation of observations within sites. This analysis yielded qualitatively similar results (Supplement). We also obtained very similar results in additional sensitivity analyses in which we fit a model to only the subset of data from similar sites (dropping the MTurk coefficient) or in which we fit meta-analytic counterparts to the primary model (Supplement).

Table 3: In units of perceived likelihood on a 0-10 scale, estimates of the main effect and focus interaction effect in similar university sites and under the RPP protocol (MTurk), as well as estimates of the difference between these estimates. Total n = 3727.

Parameter	Estimate	95% CI	p-value
Tempt main effect within MTurk	0.21	[-0.01, 0.43]	0.06
Tempt main effect within similar sites	0.11	[-0.34, 0.56]	0.64
Effect of similar site vs. MTurk on tempt main effect	-0.11	[-0.61, 0.39]	0.68
Tempt-load interaction within MTurk	-0.20	[-0.53, 0.13]	0.24
Tempt-load interaction within similar sites	-0.02	[-0.68, 0.63]	0.94
Effect of similar site vs. MTurk on tempt-load interaction	0.18	[-0.56, 0.91]	0.64

interpretations for each coefficient of interest are provided in the preregistered protocol.

The primary analysis model included 3727 subjects from similar sites and MTurk. Consistent with the RPP replication, the present results collected on MTurk did not strongly support the main effect of tempting fate (Table 3), and nor did results collected under the updated protocol in similar sites (Table 3, row 2). Updating the protocol did not appear to change the main effect estimate (Table 3, row 3). Furthermore, results from the new MTurk sample also did not support the focus interaction (Table 3, row 4), and nor did results under the updated protocol (Table 3, row 5). As seen for the main effect, updating the protocol did not meaningfully affect the focus interaction estimate (Table 3, row 6). Both the main effect of tempting fate and the focus interaction appeared homogeneous across sites (estimated random intercept standard deviation = 0; estimated random slope standard deviation = 0).

Secondary analyses: All university sites

Planned secondary analyses addressed the same questions as the primary analyses, but additionally incorporated data from dissimilar university sites (total $n = 4441$). 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 similar and dissimilar sites. Results (Table 4) did not support the main effect or the focus interaction in any site type. The main effect estimate in dissimilar sites was comparable to that in similar sites (Table 4, row 4), as was the interaction estimate (Table 4, row 8). The main effect and interaction appeared more heterogeneous across sites than in primary analyses (estimated random intercept standard deviation = 0.19; estimated random slope standard deviation = 0.37). We conducted post hoc secondary analyses (Supplement) to assess the extent to which the replication findings were statistically consistent with the original study; that is, whether it is plausible that the original study was drawn from the same distribution as the replications (Mathur & VanderWeele, 2017).

Evaluating proposed explanations for replication failure

Anticipating that results may have differed between similar and dissimilar sites, we planned to conduct secondary analyses assessing proposed explanations for the previous replication failure in RPP. However, given that results did not appear to differ between similar and dissimilar sites, we decided post hoc to pursue the following simplified secondary analyses. First, it is possible that the cognitive load manipulation could not be implemented reliably in an online setting due, for example, to 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,

[5]An alternative model specification in which all universities were treated as a single category yielded similar results (Supplement).

Table 4: In units of perceived likelihood on a 0-10 scale, estimates of the main effect and focus interaction effect in similar university sites, dissimilar university sites, and under the RPP protocol (MTurk), as well as estimates of the difference between these estimates. Total $n = 4441$.

Parameter	Estimate	95% CI	p-value
Tempt main effect within MTurk	0.21	[-0.22, 0.65]	0.34
Tempt main effect within similar sites	0.08	[-0.40, 0.57]	0.73
Tempt main effect within dissimilar sites	0.42	[-0.07, 0.90]	0.09
Effect of similar vs. dissimilar site on tempt main effect	-0.33	[-1.02, 0.36]	0.35
Tempt-load interaction within MTurk	-0.20	[-1.00, 0.60]	0.62
Tempt-load interaction within similar sites	0.01	[-0.73, 0.76]	0.97
Tempt-load interaction within dissimilar sites	-0.28	[-1.01, 0.45]	0.45
Effect of similar vs. dissimilar site on tempt-load interaction	0.29	[-0.75, 1.34]	0.58

cognitive load, and an indicator for whether a subject completed the experiment on MTurk or at any university. The three-way interaction estimate suggested that the magnitude of the focus interaction – that is, the strength of influence of the cognitive load manipulation on the tempting-fate effect – was nearly identical for MTurk subjects versus university subjects (modeled $n = 4441$; $b =$ -0.03 with 95% CI: [-0.99, 0.93]; $p = 0.95$).

To assess the effectiveness of the cognitive load manipulations, we used subjects[6] assigned to cognitive load to fit separate linear mixed models regressing perceived effort (modeled $n = 1852$) and perceived difficulty ($n = 1848$) on an indicator for whether a subject was recruited on MTurk or from any university. If, as hypothesized, the cognitive load manipulation was less effective on MTurk than in university settings, perceived effort or

[6] Due to an error in data collection, the new measures for perceived effort and difficulty were omitted for one site (University of California at Berkeley); thus, these subjects were excluded in these analyses.

difficulty might be lower for MTurk subjects. In contrast, perceived effort associated with the cognitive load task was comparable for MTurk and university subjects ($b = 0.63$ with 95% CI: [-0.42, 1.68]; $p = 0.24$), as was perceived difficulty ($b = 0.51$ with 95% CI: [-0.11, 1.14]; $p = 0.11$). Ultimately, these results do not suggest reduced efficacy of the cognitive load manipulation when implemented online versus in person.

To assess differences in academic attitudes, we used subjects[7] from all types of sites, including MTurk, to fit linear mixed models regressing perceived importance ($n = 4175$) and perceived negativity ($n = 4172$) on site type (similar, dissimilar, or MTurk) with random intercepts by site. Contrary to our speculation, MTurk subjects reported, if anything, that answering questions correctly was somewhat more important than did subjects at similar universities ($b = 1.02$ with 95% CI: [0.45, 1.59]; $p = 4.60 \times 10^{-4}$) or at dissimilar universities ($b = 0.76$ with 95% CI: [0.24, 1.29]; $p = 4.60 \times 10^{-3}$). Additionally, when asked to assess how bad it would be to answer incorrectly, MTurk subjects responded comparably to subjects at similar sites ($b = -0.03$ with 95% CI: [-0.52, 0.45]; $p = 0.89$) and at dissimilar sites ($b = 0.45$ with 95% CI: [0.01, 0.90]; $p = 0.05$).

Lastly, in a planned analysis, we assessed variation in results according to a site's similarity to Cornell, now redefining similarity using a continuous proxy (namely, a university's estimated median total SAT score in 2018) rather than the dichotomous "similar" versus "dissimilar" eligibility criterion for primary analyses. Subjects from universities outside the United States or from MTurk were excluded from this analysis, leaving an analyzed $n = 975$ from 7 universities with median SAT scores ranging from 1182 to 2178 of 2400 possible points. We assumed that universities with higher SAT scores would be most similar to Cornell (median SAT: 2134) and therefore considered a linear effect of median SAT score as a moderator of the main effects and interaction of tempting fate with cognitive load. A mixed model did not suggest that median SAT score moderated either the main

[7] These analyses again excluded subjects from UC Berkeley, which did not collect the new measures due to a data collection error.

effect of tempting fate ($b = 0.00$ for a 10-point increase in SAT score with 95% CI: [-0.01, 0.02]; $p = 0.83$) or the focus interaction ($b = 0.00$ with 95% CI: [-0.02, 0.02]; $p = 0.97$).

Comparison to results of Many Labs 2

As mentioned in the Introduction, an independent multisite replication of this experiment (Many Labs 2, or "ML2") found strong evidence for a main effect of tempting fate that was in the same direction as that of the original study, but of smaller size; this finding stands in contrast to the present study's negligible main effect estimate. We corresponded with the lead author of ML2 to identify protocol differences that might explain the discrepant results. We identified minor differences in questionnaire design, sampling frame, and statistical analysis (Supplementary Table 4) and pursued post hoc, exploratory analyses to gauge whether these differences were likely to explain the discrepant results (see Supplement for details). For example, we compared results between the two studies' United States MTurk samples to identify questionnaire design effects while holding constant the subject population, we reanalyzed the raw data from both studies 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 studies' point estimates remained meaningfully different in all of 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, which suggested strong evidence for a small main effect of tempting fate (Supplement).

Conclusion

We used an updated replication 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 Amazon Mechanical Turk, as in the previous replication, as well as at less similar universities. 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 in all direct replications, our replication was limited to a single operationalization of the tempting-fate effect; our results do not necessarily generalize to other experimental scenarios, for example.

Under the updated protocol in similar sites, we estimated a negligible main effect of tempting fate in the absence of cognitive load as well as a negligible focus interaction between tempting fate and cognitive load. Results did not appear to differ between data collected under the updated protocol in similar sites and data collected under the previous replication protocol on Amazon Mechanical Turk, nor did they differ meaningfully in dissimilar universities. Secondary analyses did not support proposed mechanisms of 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 did not provide compelling evidence for statistical inconsistency between the original study and replications under the original protocol for the main effect or for the focus interaction. Ultimately, our results fail to support the tempting-fate effect and interaction and also fail to support proposed substantive mechanisms for the replication failure in Open Science Collaboration (2015). 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, which and extensive post hoc analyses were unable to identify the source of the discrepancy.

Contributions

CRE conceived the Many Labs project. MBM, CRE, and MCF designed this multisite replication study. MBM and DJBP oversaw administration. MBM planned and conducted statistical analyses (with MCF auditing the code) and wrote the manuscript. The remaining authors collected data, audited site-level analyses, and approved the final manuscript. The authors have no conflicts of interest with respect to the authorship or publication of this manuscript. All authors approved the final manuscript with one exception (sadly, SP passed away before the manuscript draft was written).

Acknowledgments

We are grateful to Jane Risen and Thomas Gilovich for their thoughtful and responsive feedback through the protocol design process. 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.

Funding

This research was supported by a grant from the Association for Psychological Science. MHB was supported by F31AA024358. GN was supported by the Wharton Behavioral Lab.

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*(2), 328.

Klein, R. et a. (2017). Many labs 2: Investigating variation in replicability across sample and setting. *Preprint provided by authors.*

Mathur, M., & Frank, M. (2012). Replication of "Why people are reluctant to tempt fate" by Risen & Gilovich. retrieved from https://osf.io/nwua6/.

Mathur, M., & VanderWeele, T. (2017). New statistical metrics for multisite replications. *Preprint retrieved from https://osf.io/w89s5/.*

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 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*(2), 293.

Supplementary Analyses for: "Registered Multisite Replication of Tempting-Fate Effects in Risen & Gilovich (2008)"

Contents

Descriptive Statistics and Plots

Cohen's d and t-tests within each site

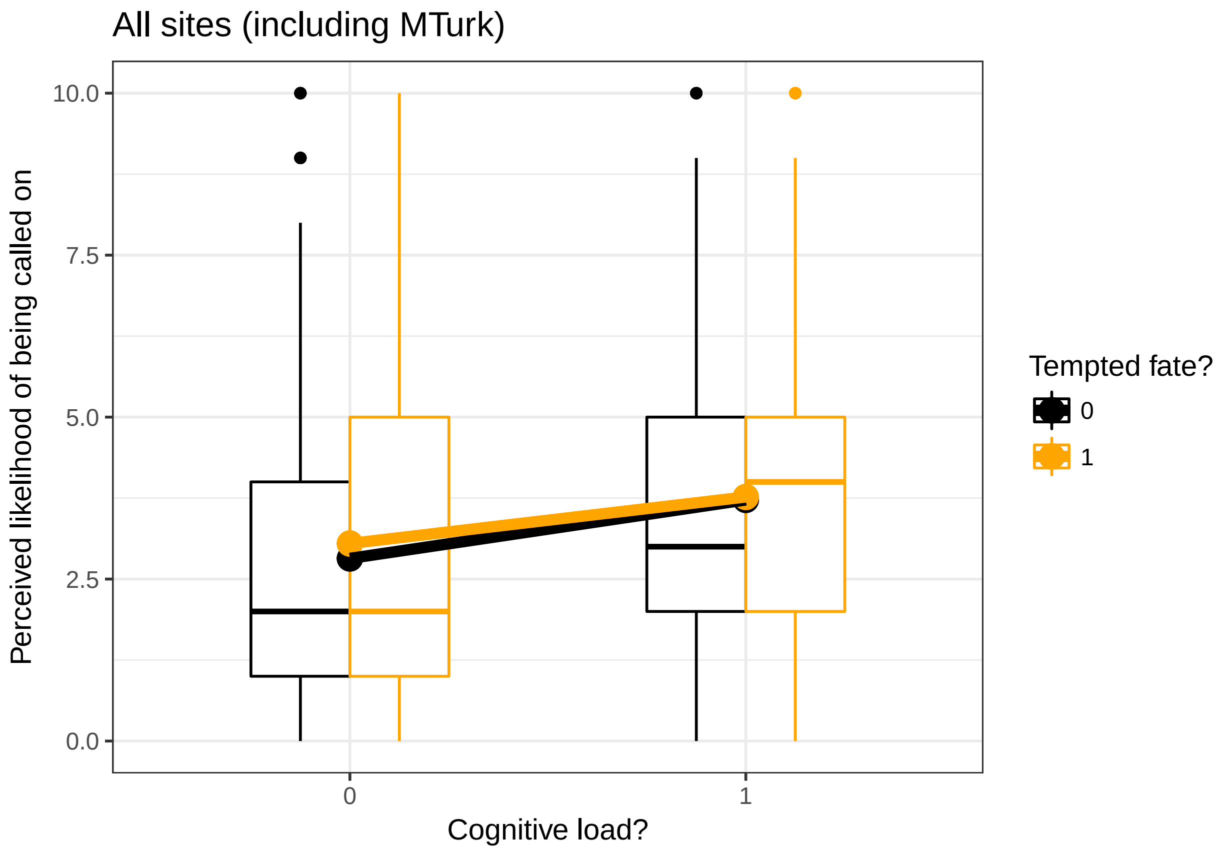
Per the preregistration, here we conduct additional within-site analyses that reproduce the original study's stratified analyses and effect sizes of tempting fate.

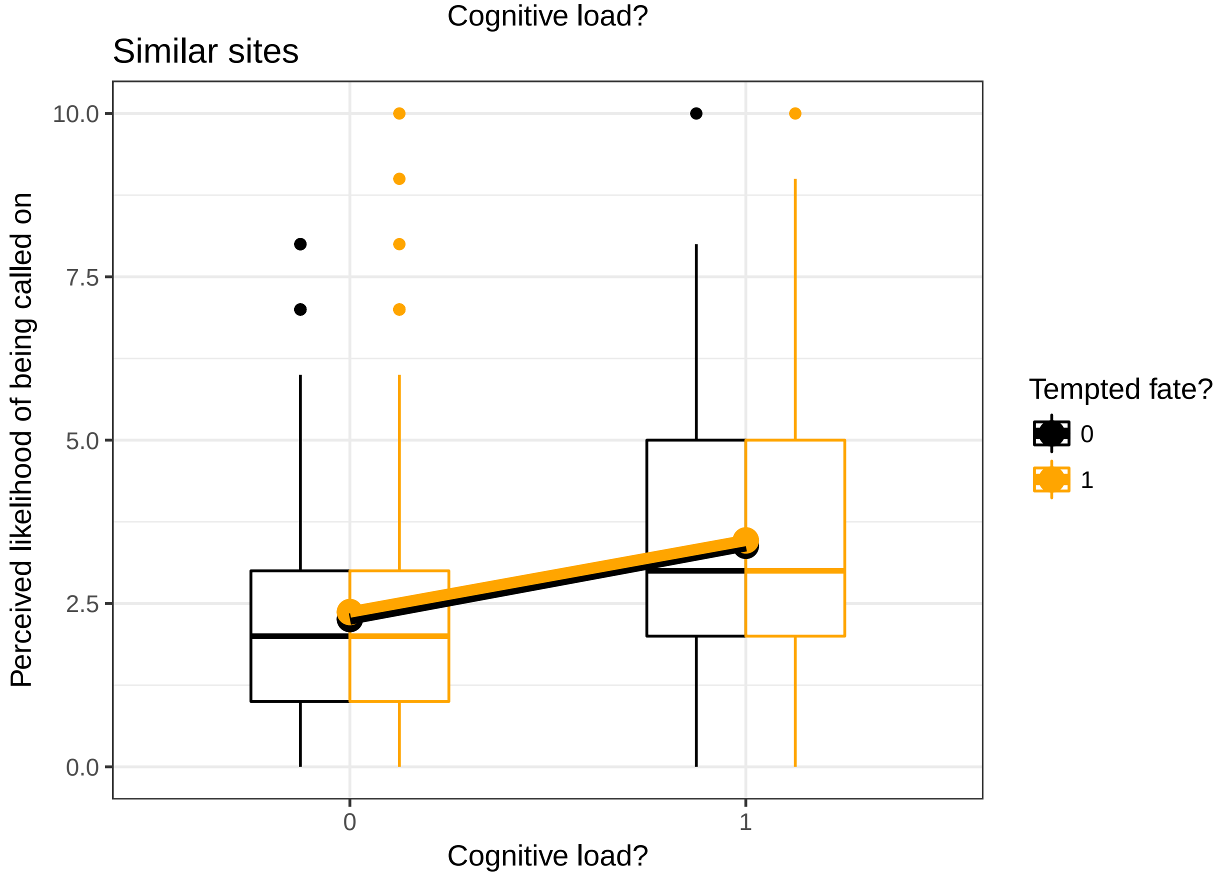
Supplementary Table 1: Cohen's d and p-values for t-tests of the effect of tempting fate on perceived likelihood, stratified by cognitive load within each site.

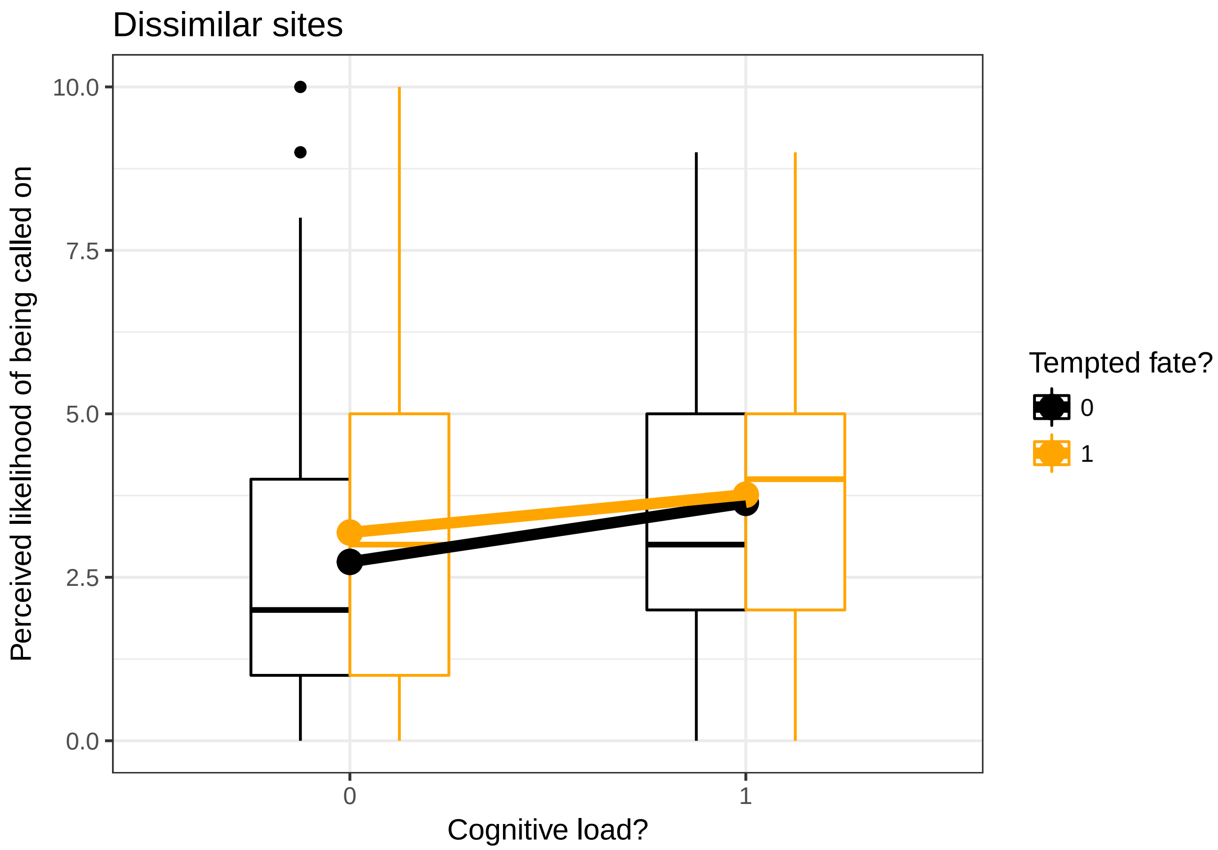
Site	Cohen's d (no load)	p-value (no load)	Cohen's d (load)	p-value (load)
MTurk	0.09	0.07	0.01	0.92
U Penn	0.13	0.40	0.20	0.20
UCB	-0.11	0.55	-0.30	0.16
UVA	0.02	0.94	0.06	0.79
Stanford	0.40	0.22	0.20	0.59
Eotvos	0.50	0.00	0.03	0.84
KUL	0.06	0.81	0.58	0.03
UP	0.00	1.00	0.02	0.95
BYUI	0.04	0.90	-0.10	0.76
URI	0.03	0.91	0.02	0.95
RHIT	0.10	0.79	-0.02	0.97

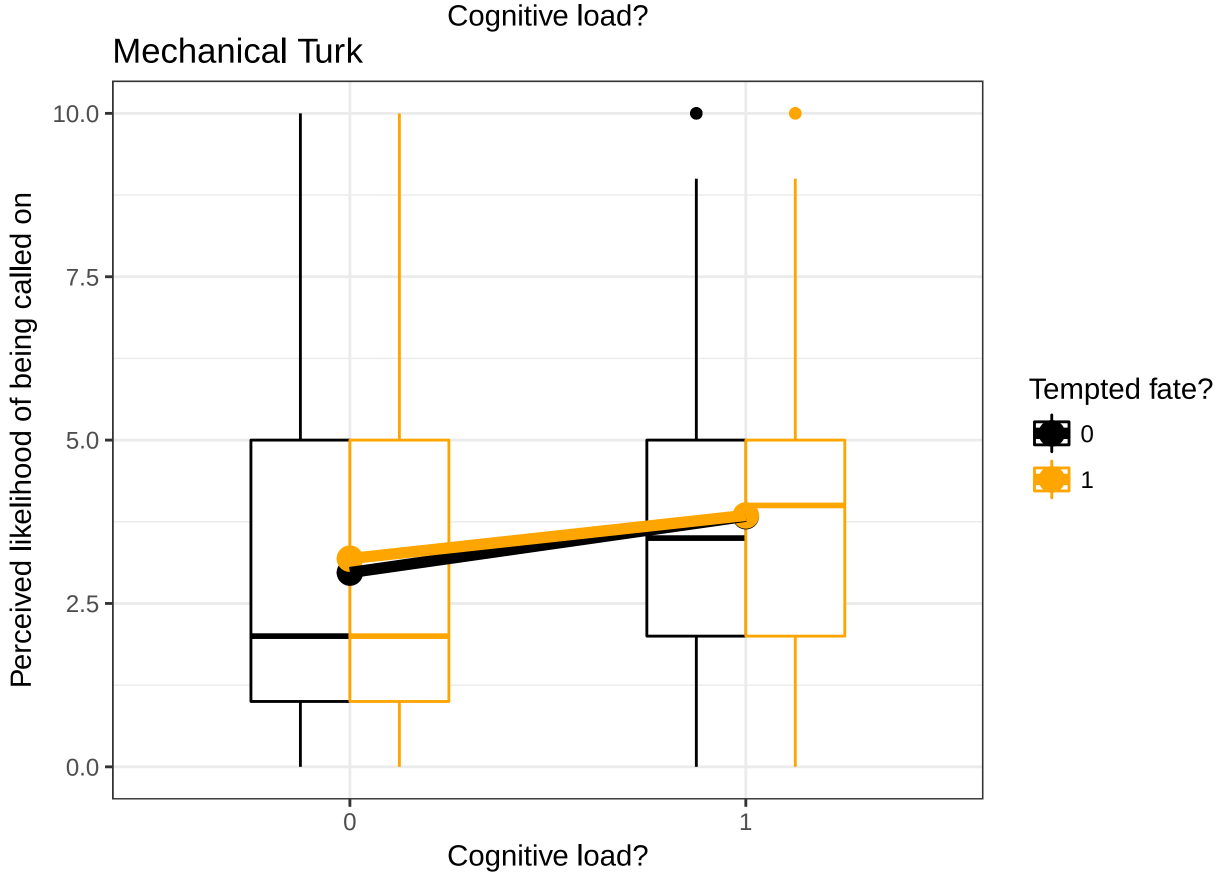
Interaction plots by site type

Boxplots: medians and IQRs; lines: simple means by subset. These aggregated means and SDs pool across all sites within a group (similar, dissimilar, MTurk) and do not account for clustering by site.


All sites (including MTurk)
Perceived likelihood of being called on
10.0
7.5
5.0
2.5
0.0
0
1
Cognitive load?
Tempted fate?
0
1


Similar sites
Perceived likelihood of being called on
10.0
7.5
5.0
2.5
0.0
0
1
Cognitive load?
Tempted fate?
0
1


Dissimilar sites
Perceived likelihood of being called on
10.0
7.5
5.0
2.5
0.0
0
1
Cognitive load?
Tempted fate?
0
1


Mechanical Turk
Perceived likelihood of being called on
10.0
7.5
5.0
2.5
0.0
0
1
Cognitive load?
Tempted fate?
0
1

Cell means and standard deviations by site type

Supplementary Table 2: Means and SDs of perceived likelihood across all subjects within each site type (naively pooling all sites)

tempt	load	group	mean	SD
0	0	a.mturk	2.97	2.39
1	0	a.mturk	3.19	2.41
0	1	a.mturk	3.84	2.31
1	1	a.mturk	3.85	2.32
0	0	b.similar	2.26	1.89
1	0	b.similar	2.37	1.96
0	1	b.similar	3.38	2.16
1	1	b.similar	3.47	2.19
0	0	c.dissimilar	2.74	2.02
1	0	c.dissimilar	3.18	2.19
0	1	c.dissimilar	3.64	2.10
1	1	c.dissimilar	3.76	2.01

Statistical Consistency of Original Study with Replications

We conducted post hoc secondary analyses to assess the extent to which the replication findings were statistically consistent with the original study; that is, whether it is plausible that the original study was drawn from the same distribution as the replications (Mathur and VanderWeele 2017). These analyses account for uncertainty in both the original study and the replication and for possible heterogeneity in the replications, and they can help distinguish whether an estimated effect size in the replications that appears to disagree with the original estimate may nevertheless be statistically consistent with the original study due, for example, to low power in the original study or in the replications or to heterogeneity.

We found that, if indeed the original study were statistically consistent with results from the similar sites in the sense of being drawn from the estimated distribution of the replications in similar sites, there would be a probability of $P_{orig} = 0.12$ that the original main effect estimate would have been as extreme as or more extreme than the observed value of $b = 1.03$. This probability is slightly higher (0.18) when considering the estimated distribution in all university sites. For the focus interaction, the probability of an original estimate at least as extreme as the observed $b = 1.54$ if the original study were statistically consistent with the similar-site replications is $P_{orig} = 0.07$; this probability is comparable (0.05) when considering the distribution of all university sites.

Sensitivity Analyses for Reported Results

Fit subset model counterpart to primary analysis model

Instead of fitting a model that includes both MTurk and similar sites with an interaction of site type, we fit a model to only the subset of similar sites.

```
m1.temp = lmer(lkl ~ tempt * load + (tempt * load | site), data = b[b$group ==
    "b.similar", ])
CI.temp = confint(m1.temp, method = "Wald")
```

In the primary model, the estimated main effect was 0.11 with 95% CI: (-0.34, 0.56), whereas in the present subset model, it is 0.13 with 95% CI: (-0.33, 0.60). Also, in the primary model, the estimated interaction effect was -0.02 with 95% CI: (-0.68, 0.63), whereas in the present subset model, it is -0.03 with 95% CI: (-0.66, 0.6). These results are similar.

Fit meta-analytic counterparts to primary analysis model

Instead of fitting a mixed model to observation-level data, we fit a random-effects meta-analysis to the point estimates using the Paule & Mandel heterogeneity estimator and the Knapp-Hartung standard error adjustment. For the main effect:

```
meta.main = rma.uni(yi = site.main.est, vi = site.main.SE^2,
    data = sites[sites$group == "b.similar", ], measure = "MD",
    method = "PM", knha = TRUE)

p.orig.main.2 = p_orig(orig.y = yi.orig.main, orig.vy = vyi.orig.main,
    yr = meta.main$b, t2 = meta.main$tau2, vyr = meta.main$vb)
```

In the above mixed model, the estimated main effect and heterogeneity in similar sites was $\widehat{M} = 0.13$ and $\widehat{V} = 0.06$ compared to $\widehat{M} = 0.13$ and $\widehat{V} = 0$ in the meta-analysis. They agree very closely. P_{orig} is a bit lower (0.07) due to the lower estimated heterogeneity here.

For the focus interaction effect:

```
meta.int = rma.uni(yi = site.int.est, vi = site.int.SE^2, data = sites[sites$group ==
    "b.similar", ], measure = "MD", method = "PM", knha = TRUE)

p.orig.int.2 = p_orig(orig.y = yi.orig.int, orig.vy = vyi.orig.int,
    yr = meta.int$b, t2 = meta.int$tau2, vyr = meta.int$vb)
```

In the above mixed model, the estimated interaction effect and heterogeneity in similar sites was $\widehat{M} =$ -0.03 and $\widehat{V} = 0.06$ in the mixed model compared to $\widehat{M} =$ -0.02 and $\widehat{V} = 0$ in the meta-analysis. P_{orig} is again slightly lower (0.04). These results agree reasonably closely.

Combine all universities into one category

In the planned secondary analysis model including all universities, similar and dissimilar sites were treated as separate categories. Here, they are combined into one category.

Supplementary Table 3: Main effect and interaction estimates when combining all universities

Name	Estimate	CI	pval
Tempt main effect within MTurk	0.21	[-0.28, 0.71]	0.40

Name	Estimate	CI	pval
Tempt main effect within university sites	0.28	[-0.08, 0.63]	0.12
Effect of university site vs. MTurk on tempt main effect	0.06	[-0.55, 0.67]	0.84
Tempt-load interaction within MTurk	-0.20	[-1.01, 0.60]	0.62
Tempt-load interaction within university sites	-0.17	[-0.69, 0.35]	0.52
Effect of university site vs. MTurk on tempt-load interaction	0.03	[-0.93, 0.99]	0.95

Refit original study's ANOVA model

The original study used two-way ANOVA to test for the main effect and interaction. Per our preregistered protocol, we also reproduce this model as a secondary analysis here. Since this model is statistically equivalent to the regression models presented in the main text, this is simply a different way of presenting the contrasts. The results are qualitatively similar to those in the main text.

```
# with standard ANOVA mean contrasts and sequential
# decomposition main effect: half the effect of tempting fate
# vs. not tempting fate when not under load
summary(aov(lkl ~ tempt * load, data = b[b$group == "b.similar",
    ]))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## tempt          1    1.4    1.36   0.325    0.569
## load           1  230.5  230.53  55.009 3.25e-13 ***
## tempt:load     1    0.0    0.03   0.007    0.933
## Residuals    750 3143.0    4.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# with contrasts vs. 0 and marginal SS decomposition main
# effect: effect of tempting fate when not under load
summary(lm(lkl ~ tempt * load, data = b[b$group == "b.similar",
    ]))
```

```
##
## Call:
## lm(formula = lkl ~ tempt * load, data = b[b$group == "b.similar",
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4655 -1.3829 -0.3829  1.5345  7.6311
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.26131    0.14512  15.583  < 2e-16 ***
## tempt         0.10763    0.20347   0.529    0.597
## load          1.12155    0.21215   5.287 1.63e-07 ***
## tempt:load   -0.02497    0.29905  -0.083    0.933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.047 on 750 degrees of freedom
## Multiple R-squared:  0.06872,    Adjusted R-squared:  0.06499
## F-statistic: 18.45 on 3 and 750 DF,  p-value: 1.485e-11
```

Characteristic	Original protocol	ML2	ML5	Testable implication	Conclusion
Dimensions of scale measuring perceived likelihood	Likelihood scale contained 11 options, ranging from 0 to 10.	Likelihood scale contained 10 options, ranging from 1 to 10.	Likelihood scale contained 11 options, ranging from 0 to 10.	Questionnaire design effects may produce differences in results between the otherwise comparable U.S. Mturk samples.	Unlikely; no meaningful differences in Mturk samples.
Missing data	Unknown, but likely allowed subjects to skip questions due to the pencil-and-paper format.	The online questionnaire allowed subjects to skip questions and experiments.	The online questionnaire required subjects to answer all questions.	Questionnaire design effects may produce differences in results between the otherwise comparable U.S. Mturk samples.	Unlikely; no meaningful differences in Mturk results and proportion of missing data <1% in ML2.
Presence of unrelated experiments	Experiment was administered alone.	Experiment was administered as part of a 30-minute block of experiments, with order randomized.	Experiment was administered alone, or only after other experiments pre-approved as unlikely to influence results.	If other tasks interfered in ML2, results may differ by the order of presentation.	Unlikely; order effects appeared minimal and non-systematic in ML2.
Sampling frame for primary analyses	Undergraduates at Cornell University	Undergraduates	Undergraduates at "similar" sites	Comparing all ML2 sites to all ML5 sites may result in better agreement, since these sampling frames are more directly comparable.	Unlikely; results still discrepant.
Statistical analysis	One-way ANOVA (assuming homoskedasticity)	Independent-samples t-test combining sites' data (not assuming homoskedasticity)	Linear mixed model combining sites' data and using model-based SEs (assuming homoskedasticity)	Aggregating individual subject level data from ML2 and ML5 using a single analysis method may result in better agreement.	Unlikely; results still discrepant using LMM, GEE, and RMD analysis approaches.
Outlying sites	Outlying sites	None evident	A single site with a small sample size estimated a large, positive main effect.	N/A	Unlikely; the only possible outlier would have made results less discrepant.

Supplementary Table 4: Comparison of experimental protocols used in the original study, the RPP replication, and the present replication.

Comparison to Many Labs 2

Questionnaire differences

We identified three minor differences in the design of the questionnaire that might have contributed to the discrepant results (Supplementary Table 4, rows 1-3). First, the endpoints of the scale measuring perceived likelihood ranged from 1-10 in ML2 but ranged from 0-10 in ML5 and in the original study. Second, the questionnaire permitted subjects to skip questions in ML2, but not in ML5, potentially leading to systematic differences in characteristics of self-selected subjects; the original study was conducted on paper, so likely did not prevent subjects from skipping questions. Third, in ML2, the questionnaire was embedded in a roughly 30-minute series of experiments presented in a randomized order, which was not the case in ML5.

To investigate whether these differences in questionnaire design might have contributed to the discrepant results, we compared results between each study's sample of Amazon Mechanical Turk subjects in the United States. To the extent that these samples are directly comparable, any effects of the different questionnaire designs would likely produce differences in results between the MTurk samples. In contrast, the two samples estimated nearly identical point estimates (raw mean difference = 0.21, n = 340 in ML2 vs. 0.21, n = 2973 in ML5)[1]. Additionally, the extent of missing data was negligible in ML2 (0.5%), and their analyses suggested that results for this experiment differed little based on the order in which the experiment was presented relative to the unrelated experiments (Klein 2017).

Analysis differences

Additionally, the two studies used slightly different statistical analyses to aggregate data across sites: ML2 used an independent-samples t-test allowing for heteroskedasticity, while ML5 used a linear mixed model with model-based standard errors that assumed homoskedasticity. The original study used a one-way

[1]Of course, statistical inference on these point estimates differed due to the difference in sample sizes.

ANOVA (which assumes homoskedasticity). We therefore reanalyzed all subject-level data for primary analysis sites in both studies using identical statistical analyses using three approaches[2]:

1. To reproduce the analysis approach in ML5, we used a linear mixed model (**LMM**) to regress perceived likelihood on fixed effects of tempting fate, the interaction of tempting fate with study (ML5 vs. ML2), exchangeable random intercepts and slopes by site, and with normal errors. Inference used model-based standard errors, so assumed homoskedasticity.

2. To avoid parametric assumptions[3], we used generalized estimating equations (**GEE**) with a working independent correlation structure to regress perceived likelihood on tempting fate. Inference used robust standard errors, so made no assumptions about how subjects were correlated within sites or otherwise.

3. To reproduce the analysis approach in ML2, we computed raw mean differences (**RMD**) between tempting-fate conditions for each study separately, ignoring correlation of subjects within sites. We estimated standard errors for each as in the independent-samples Welch t-test conducted for ML2. We then tested the null hypothesis that the RMDs reflected the same population difference in both studies using the fact that, under the null:

$$\frac{\widehat{\Delta}_{ML2} - \widehat{\Delta}_{ML5}}{\sqrt{\widehat{SE}^2_{\widehat{\Delta}_{ML2}} + \widehat{SE}^2_{\widehat{\Delta}_{ML5}}}} \approx N(0,1)$$

   where $\widehat{\Delta}_{ML2}$ and $\widehat{\Delta}_{ML5}$ denote the estimated RMDs in ML2 and ML5 respectively.

All three approaches yielded strong evidence for a smaller average main effect size in ML5 versus ML2 (Supplementary Table 5).

Supplementary Table 5: Difference between ML5 and ML2 in average main effect estimates (as raw mean differences) in primary analysis sites under varying statistical assumptions

Method	Allows correlation within sites	Assumes homoskedasticity	Estimate	95% CI	p-value
LMM	Yes, with normal site effects	Yes	-0.45	[-0.82, -0.08]	0.02
GEE	Yes, without assumptions on structure	No	-0.45	[-0.83, -0.07]	0.02
RMD	No	Yes	-0.45	[-0.78, -0.12]	0.01

Sampling frame differences

The most conspicuous protocol difference involved the sampling frame used in primary analyses: ML2 analyzed undergraduates at a variety of colleges and universities in the United States and abroad, while ML2 analyzed undergraduates at similar sites. We speculated that results might align more closely if we constructed more directly comparable sampling frames, so we redid all three analyses above (LMM, GEE, and RMD), but expanding the sampling frame to all sites that collected data for each study. Thus, the expanded sampling frame for each study included subjects collected in one or more online samples, undergraduates at a small number of universities that would be classified as "similar" by ML5, and undergraduates at a larger number of domestic or foreign universities that would be classified as "dissimilar" by ML5.

[2]Each of these models is saturated, so is unbiased for the point estimate; thus, differences would emerge primarily in statistical inference.

[3]Although GEE is in general semiparametric rather than nonparametric, here the model is saturated due to the categorical predictors, so is effectively nonparametric.

Supplementary Table 6: Difference between ML5 and ML2 in average main effect estimates (as raw mean differences) in all sites under varying statistical assumptions

Method	Allows correlation within sites	Assumes homoskedasticity	Estimate	95% CI	p-value
LMM	Yes, with normal site effects	Yes	-0.27	[-0.52, -0.02]	0.03
GEE	Yes, without assumptions on structure	No	-0.29	[-0.46, -0.13]	0.00
RMD	No	Yes	-0.29	[-0.47, -0.12]	0.00

Lastly, we investigated whether outlying sites might have strongly influenced results in one or both studies. These outliers could arise, for example, from idiosyncrasies of protocol administration, subject characteristics, or errors in data collection or analysis. From a visual inspection of the forest plots, none of the primary analysis sites in either study appeared to be an outlier. Among all sites, Eotvos Lorand University (in ML5) may have been a modest outlier, but because this site estimated a large positive effect, its presence would, if anything, have reduced rather than exacerbated the discrepancy between studies. Outlying sites therefore do not appear to account for the discrepancy.

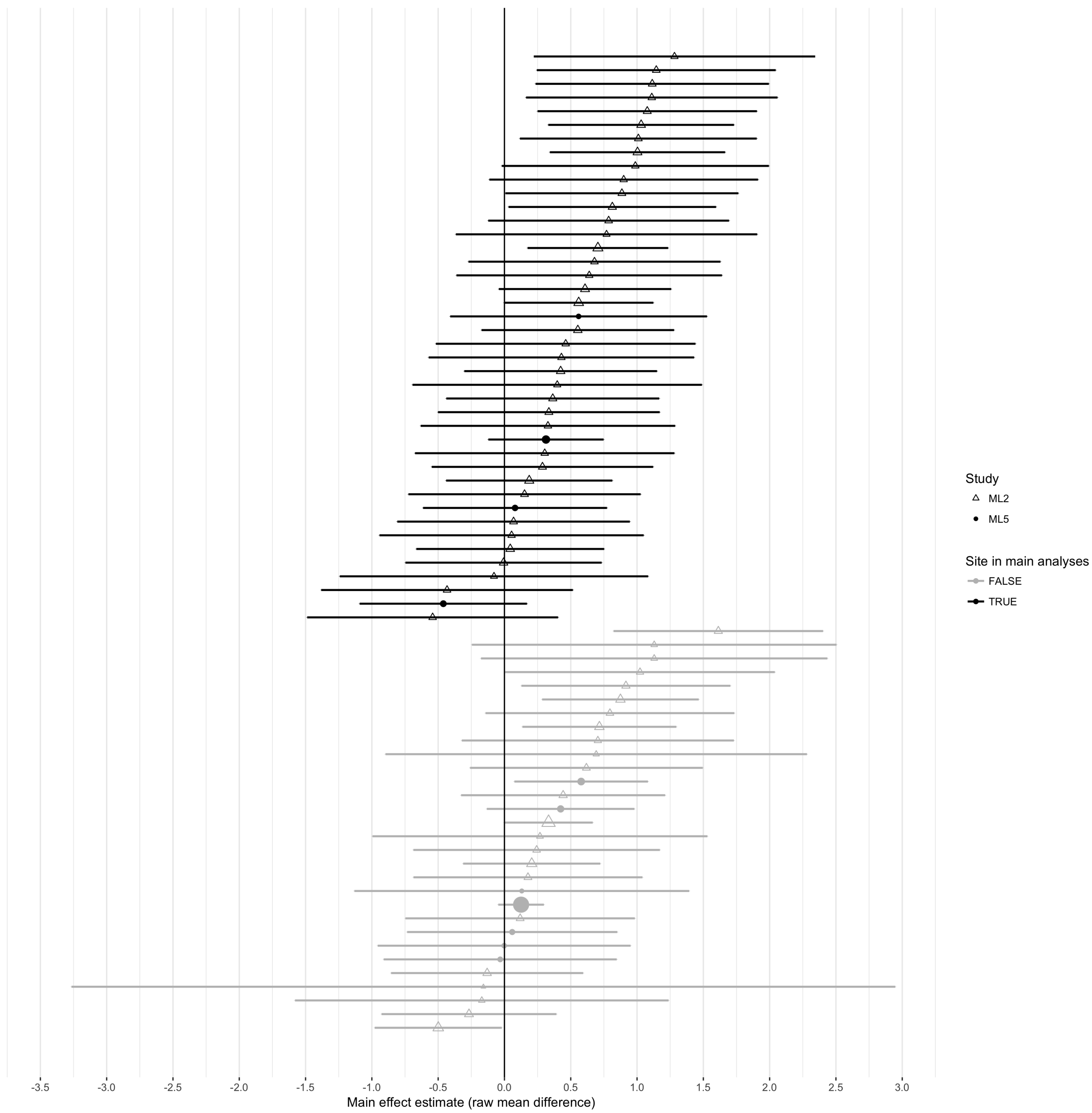
Combined estimates of tempting-fate effects

We combined all data from both studies to arrive at an updated estimate of the average effect of tempting fate across the diverse samples included in both studies. We again fit LMM and GEE models to primary analysis sites from both studies, omitting the interaction of tempting fate with study to estimate an average effect. The LMM model estimated a tempting-fate effect of 0.48 on the raw mean difference scale (95% CI: [0.34, 0.61]; $p = 8 \times 10^{-12}$). As expected, the GEE yielded a similar estimate (0.46) with more conservative inference (95% CI: [0.32, 0.61]; $p = 4 \times 10^{-10}$). The naïve Cohen's d effect size was an estimated 0.19 (95% CI 0.14, 0.25). To assess consistency between this pooled estimate and the estimate of the original study, we estimated that, if the original study were consistent with all replications, then the probability of observing a point estimate in the original study at least as extreme as that actually observed would be approximately $P_{orig} = 0.28$ (Mathur and VanderWeele 2017). This method assumes normally distributed true effects, which appeared reasonable here.

Alternatively, including all sites that collected data for either study had little effect on point estimates or inference. The LMM model estimated a tempting-fate effect of 0.41 on the raw mean difference scale (95% CI: [0.3, 0.52]; $p = 4 \times 10^{-13}$). The GEE estimated a tempting-fate effect of 0.33 (95% CI: [0.19, 0.46]; $p = 3 \times 10^{-6}$). The naïve Cohen's d effect size was an estimated 0.13 (95% CI 0.1, 0.17). As a metric of consistency with the original study, we estimated $P_{orig} = 0.25$; however, this must be interpreted cautiously in light of possible departure from normality.

Discussion

These analyses suggest that none of the known differences in questionnaire design, statistical analysis, or sampling frame appeared to adequately explain the discrepancy in results. Combining data from both studies provided strong evidence for small effects of tempting fate, and these findings were robust to different statistical assumptions. The resulting point estimates were considerably smaller than that of the original study, but appeared statistically consistent with the original due to the latter's limited sample size.



Supplementary Figure 2: Main effect estimates for sites used in primary analyses (black) and those not used in main analysis (gray) for ML2 (triangles) and ML5 (circles). Plot symbol size is inversely proportional the estimated within-study variance.

References

Klein, R et al. 2017. "Many Labs 2: Investigating Variation in Replicability Across Sample and Setting." *Preprint provided by authors.*

Mathur, MB, and TJ VanderWeele. 2017. "New Statistical Metrics for Multisite Replications." *Preprint retrieved from https://osf.io/w89s5/.*