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

Journal:	*Advances in Methods and Practices in Psychological Science*
Manuscript ID	AMPPS-17-0118.R3
Manuscript Type:	Registered Replication Report
Date Submitted by the Author:	n/a
Complete List of Authors:	Mathur, Maya; Harvard University , Biostatistics
Method and Stats :	reproducbility
Substance Keywords:	
Additional Keywords:	replication, reproducibility, preregistered, open data, magical thinking

Running head: MULTISITE REPLICATION OF TEMPTING FATE   1

[1] Many Labs 5: Registered multisite replication of tempting-fate effects in Risen & Gilovich

[2] (2008)

[3] Maya B. Mathur[1,2], Diane-Jo Bart-Plange[3], Balazs Aczel[4], Michael H. Bernstein[5], Antonia

[4] Ciunci[5], Charles R. Ebersole[3], Filipe Falcão[6], Kayla Gerken[7], Rias A. Hilliard[7], Alan Jern[7],

[5] Danielle Kellier[2], Grecia Kessinger[8], Vanessa Kolb[5], Marton Kovacs[4], Caio Lage[9], Eleanor V.

[6] Langford[3], Samuel Lins[6], Dylan Manfredi[10], Venus Meyet[8], Don A. Moore[11], Gideon Nave[10],

[7] Christian Nunnally[7], Anna Palinkas[4], Kimberly P. Parks[3], Sebastiaan Pessers[12], Tiago

[8] Ramos[6], Kaylis Hase Rudy[8], Janos Salamon[4], Rachel L. Shubella[7], Rúben Silva[6], Sara

[9] Steegen[12], L.A.R. Stein[5,13,14], Barnabas Szaszi[4], Peter Szecsi[4], Francis Tuerlinckx[12], Wolf

[10] Vanpaemel[12], Maria Vlachou[12], Bradford J. Wiggins[8], David Zealley[8], Mark Zrubka[4], &

[11] Michael C. Frank[2]

[12] [1] Harvard University, Boston, MA, United States

[13] [2] Stanford University, Stanford, CA, United States

[14] [3] University of Virginia, Charlottesville, VA, United States

[15] [4] ELTE Eötvös Loránd University, Budapest, Hungary

[16] [5] University of Rhode Island, Kingston, RI, United States

[17] [6] University of Porto, Porto, Portugal

[18] [7] Rose-Hulman Institute of Technology, Terre Haute, IN, United States

[19] [8] Brigham Young University - Idaho, Rexburg, ID, United States

[20] [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

MULTISITE REPLICATION OF TEMPTING FATE 3

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).

82 fate vs. not tempting fate for subjects under cognitive load vs. not under cognitive load: $b =$

83 1.54 with 95% CI: [0.05, 3.03]; $p = 0.04$).

84   We selected Risen and Gilovich (2008) for replication because, per the selection criteria

85 of all Many Labs 5 replications, this study was subject to a previous replication attempt by

86 Open Science Collaboration (2015). The previous replication found little evidence for either

87 a main effect of tempting fate without cognitive load ($n = 226$, $b = 0.20$ with 95% CI: [-0.58,

88 0.97]; $p = 0.62$) or the focus interaction ($b = 0.03$ with 95% CI: [-1.14, 1.20]; $p = 0.96$)

89 (Mathur & Frank, 2012). However, prior to the collection of replication data by this previous

90 replication effort (termed "RPP"), the authors of the original study expressed concerns

91 about the replication protocol. Due to feasibility constraints, the RPP replication proceeded

92 without addressing these concerns. Specifically, the replication was implemented on the

93 crowdsourcing website Amazon Mechanical Turk, a setting that could potentially compromise

94 the cognitive load manipulation if subjects were already multitasking or were distracted.

95 Additionally, the experimental scenario, which required subjects to imagine being unprepared

96 to answer questions in class, may be less personally salient to subjects not enrolled in an

97 elite university similar to Cornell University, the site of the original study. Thus, as part of

98 the Many Labs 5 project, the present multisite replication aimed to: (1) reassess replicability

99 of Risen and Gilovich (2008) using an updated protocol designed in collaboration with the

100 original authors to mitigate potential problems with the previous replication protocol; and

101 (2) formally assess the effect of updating the protocol in this manner by comparing its results

102 to newly collected results under the previous replication protocol.

103   Concurrently with the present study, an independent group (Many Labs 2) conducted

104 a multisite replication of the main effect, but not the interaction (Klein, 2017). Their

105 primary analysis sample comprised undergraduates at universities and colleges in the United

106 States and abroad ($n = 4599$). These subjects judged the likelihood of being called on to be

107 higher when they had tempted fate (mean = 4.61, SD =2.42) than when they had not

108 tempted fate (mean = 4.07, SD = 2.36; $t(4,597) = 7.57$, $p = 4.4 \times 10^{-14}$, $d = 0.22$, 95% CI

[109] [0.17, 0.28]), providing strong evidence for a main effect of tempting fate, albeit of smaller

[110] magnitude than in the original study. We discuss the results of the present study in light of

[111] these existing findings.

Disclosures

[113] The protocol, sample size criteria, exclusion criteria, and statistical analysis plan were

[114] preregistered[2] with details publicly available (https://osf.io/8y6st/ for the protocol and

[115] https://osf.io/vqd5c/ for the analyses); departures from these plans are reported in this

[116] manuscript. All data, materials, and analysis code are publicly available and documented

[117] (https://osf.io/h5a9y/). Sites obtained ethics committee approval when appropriate to their

[118] geographical location and institutional requirements, and data were collected in accordance

[119] with the Declaration of Helsinki.

Methods

[121] In consultation with the original authors, we designed a replication procotol that more

[122] closely duplicated the original design than did the RPP replication (Table 1). Primary

[123] analyses used only data from university sites located in the United States and meeting an

[124] academic criterion for similarity to the original site (Table 1, row 1); these sites are termed

[125] "similar sites". We additionally used the previous RPP replication protocol without

[126] modification to collect a new sample on Amazon Mechanical Turk ("MTurk"). Finally, we

[127] collected secondary data in several universities not meeting the SAT criterion for similarity

[128] to Cornell or located outside the United States, henceforth termed "dissimilar sites". Data

[129] from dissimilar sites were used in secondary analyses to further increase power and assess

[130] whether, as hypothesized, site similarity in fact moderates the focus effect. For sites whose

[131] 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.

MULTISITE REPLICATION OF TEMPTING FATE 7

132 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.

133 The primary statistical estimands were (1) the focus interaction within similar sites

134 and (2) the difference in this interaction between similar sites and MTurk (modeled as a

135 three-way interaction, as described below). Sample sizes were chosen to provide, in

136 aggregate, more than 80% power to detect a three-way interaction with effect size more than

137 0.75 standard deviations of perceived likelihood. Because detecting the three-way interaction

138 requires substantially larger sample sizes than detecting the focus interaction alone, this

139 choice of sample sizes also provided > 99% power to detect, within similar sites alone, a

140 focus interaction of the size reported in the original study. Each site additionally attempted

141 to reach these power criteria internally, though in many cases this was not feasible. Site-level

142 and aggregate analyses were conducted by one author (MBM), who was blinded to results

143 until all sites had completed data collection; these analyses were audited for accuracy by

144 other authors.

145 We collected four new measures, developed in discussion with the original authors, for

146 use in secondary analyses. As manipulation checks for the effectiveness of the cognitive load

147 manipulation, we asked subjects assigned to cognitive load to assess on a 0-10 scale the

148 perceived effort associated with this task (*"How much effort did the counting task require?"*)

149 and the task's difficulty (*"How difficult was the counting task?"*). Additionally, the original

150 authors speculated that the experimental scenario (regarding answering questions in class)

151 may be personally salient to subjects in an academically competitive environment similar to

152 the site of the original study, but may be less so for MTurk subjects or subjects in dissimilar

153 universities. To assess this possibility, we developed new measures in collaboration with the

154 original authors which required subjects to evaluate on a 0-10 scale the importance of

155 answering questions correctly in class (*"If you were a student in the scenario you just read*

156 *about, how important would it be for you to answer questions correctly in class?"*) and the

157 perceived negativity of answering incorrectly (*"If you were a student in the class, how bad*

158 *would you feel if you were called on by the professor, but couldn't answer the question?"*).

Results

160 **Descriptive analyses**

161   Table 2 displays sample sizes, the number of exclusions, and protocol characteristics for

162 all sites. To estimate the main effect of tempting fate and the focus interaction within each

163 site, we fit an ordinary least squares regression model of perceived likelihood on tempting

164 fate, cognitive load, and their interaction within each site. This analysis approach is

165 statistically equivalent to the ANOVA model fit in the original study while also yielding

166 coefficient estimates that are directly comparable to those estimated in primary analysis

167 models, discussed below. Figures 1 and 2, respectively, display these within-site estimates for

168 the main effect and interaction.[3]

169   Among the 4 similar sites, 3 had main effect estimates in the same direction as the

170 original study estimate, albeit of considerably smaller magnitude. Main effect estimates in

171 similar sites had p-values ranging from 0.31 to 0.94. In the MTurk sample, the main effect

172 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.

MULTISITE REPLICATION OF TEMPTING FATE

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

MULTISITE REPLICATION OF TEMPTING FATE   10

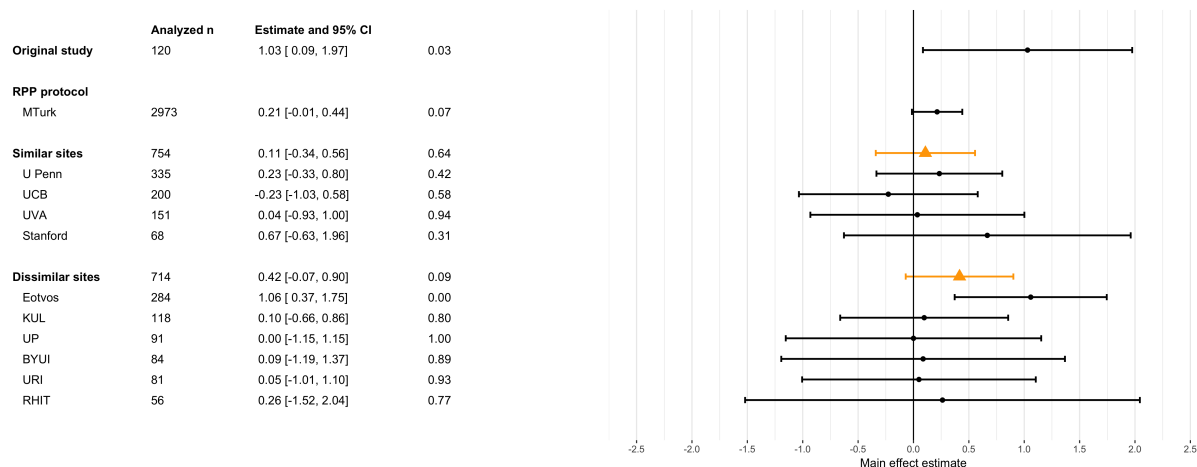|	Analyzed n	Estimate and 95% CI	
Original study	120	1.03 [0.09, 1.97]	0.03
RPP protocol			
MTurk	2973	0.21 [-0.01, 0.44]	0.07
Similar sites	754	0.11 [-0.34, 0.56]	0.64
U Penn	335	0.23 [-0.33, 0.80]	0.42
UCB	200	-0.23 [-1.03, 0.58]	0.58
UVA	151	0.04 [-0.93, 1.00]	0.94
Stanford	68	0.67 [-0.63, 1.96]	0.31
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.

181 opposite direction from the original estimate and was slightly larger in magnitude than the
182 RPP estimate. Considering all 10 university sites, 4 had point estimates in the same
183 direction as the original study, all of which were of smaller magnitude. With one exception
184 (Eotvos Lorand University with $p = 0.05$), p-values across all universities ranged from 0.21 to
185 0.99. Eotvos Lorand University obtained a large point estimate in the opposite direction
186 from the original study.

187 **Primary analyses**

188   Primary analyses aimed to: (1) estimate the focus interaction and the main effect
189 under the updated protocol in similar sites; and (2) assess whether the focus interaction and
190 the main effect estimates differed between the updated protocol and the RPP protocol. To
191 this end, we combined data from the similar sites and MTurk to fit a linear mixed model
192 with fixed effects representing main effects of tempting fate, cognitive load, and protocol

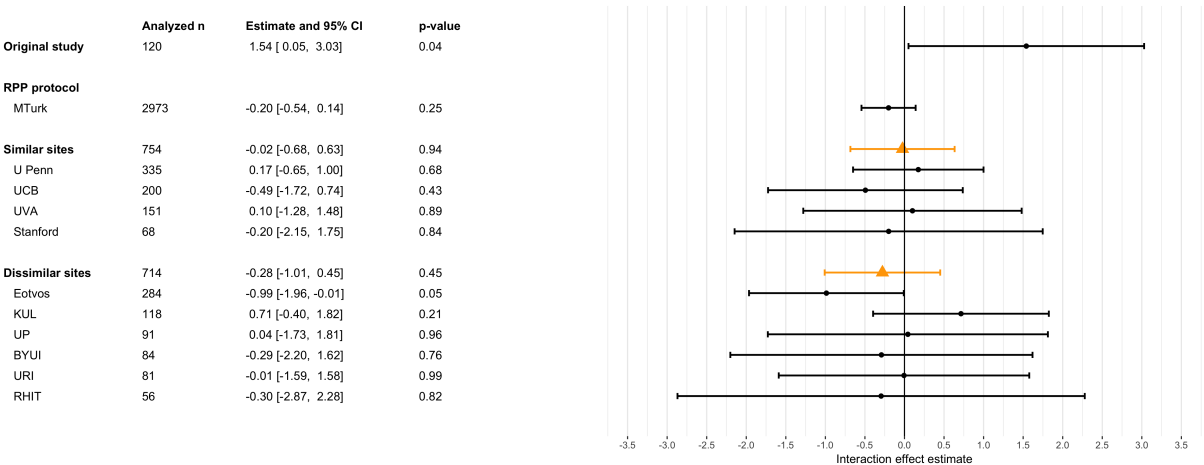|	Analyzed n	Estimate and 95% CI	p-value
Original study	120	1.54 [0.05, 3.03]	0.04
RPP protocol			
MTurk	2973	-0.20 [-0.54, 0.14]	0.25
Similar sites	754	-0.02 [-0.68, 0.63]	0.94
U Penn	335	0.17 [-0.65, 1.00]	0.68
UCB	200	-0.49 [-1.72, 0.74]	0.43
UVA	151	0.10 [-1.28, 1.48]	0.89
Stanford	68	-0.20 [-2.15, 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.

[193] (similar sites under the updated protocol vs. MTurk). To account for correlation of

[194] observations within a site, the model also contained random intercepts by site and random

[195] slopes by site of tempting fate, cognitive load, and their interaction; in all analyses, all

[196] random effects were assumed independently and identically normal.[4] This model allows

[197] estimation of the focus effect within similar sites and within MTurk and permits formal

[198] assessment of the extent to which these effects differ (via the three-way interaction of

[199] 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).

MULTISITE REPLICATION OF TEMPTING FATE   12

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

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

201   The primary analysis model included 3727 subjects from similar sites and MTurk.

202 Consistent with the RPP replication, the present results collected on MTurk did not strongly

203 support the main effect of tempting fate (Table 3), and nor did results collected under the

204 updated protocol in similar sites (Table 3, row 2). Updating the protocol did not appear to

205 change the main effect estimate (Table 3, row 3). Furthermore, results from the new MTurk

206 sample also did not support the focus interaction (Table 3, row 4), and nor did results under

207 the updated protocol (Table 3, row 5). As seen for the main effect, updating the protocol did

208 not meaningfully affect the focus interaction estimate (Table 3, row 6). Both the main effect

209 of tempting fate and the focus interaction appeared homogeneous across sites (estimated

210 random intercept standard deviation = 0; estimated random slope standard deviation = 0).

MULTISITE REPLICATION OF TEMPTING FATE   13

Secondary analyses: All university sites

211   Planned secondary analyses addressed the same questions as the primary analyses, but
213 additionally incorporated data from dissimilar university sites (total $n = 4441$). Site type
214 was treated as a categorical variable (MTurk, similar university site, or dissimilar university
215 site)[5]. Additionally, these analyses formally estimated the difference in results between
216 similar and dissimilar sites. Results (Table 4) did not support the main effect or the focus
217 interaction in any site type. The main effect estimate in dissimilar sites was comparable to
218 that in similar sites (Table 4, row 4), as was the interaction estimate (Table 4, row 8). The
219 main effect and interaction appeared more heterogeneous across sites than in primary
220 analyses (estimated random intercept standard deviation = 0.19; estimated random slope
221 standard deviation = 0.37). We conducted post hoc secondary analyses (Supplement) to
222 assess the extent to which the replication findings were statistically consistent with the
223 original study; that is, whether it is plausible that the original study was drawn from the
224 same distribution as the replications (Mathur & VanderWeele, 2017).

Evaluating proposed explanations for replication failure

226   Anticipating that results may have differed between similar and dissimilar sites, we
227 planned to conduct secondary analyses assessing proposed explanations for the previous
228 replication failure in RPP. However, given that results did not appear to differ between
229 similar and dissimilar sites, we decided post hoc to pursue the following simplified secondary
230 analyses. First, it is possible that the cognitive load manipulation could not be implemented
231 reliably in an online setting due, for example, to competing distractions in subjects'
232 uncontrolled environments (Rand, 2012). We therefore assessed the extent to which the
233 efficacy of the cognitive load manipulation differed between MTurk subjects and all
234 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).

MULTISITE REPLICATION OF TEMPTING FATE   14

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

235 cognitive load, and an indicator for whether a subject completed the experiment on MTurk

236 or at any university. The three-way interaction estimate suggested that the magnitude of the

237 focus interaction – that is, the strength of influence of the cognitive load manipulation on

238 the tempting-fate effect – was nearly identical for MTurk subjects versus university subjects

239 (modeled $n = 4441$; $b = $ -0.03 with 95% CI: [-0.99, 0.93]; $p = 0.95$).

240   To assess the effectiveness of the cognitive load manipulations, we used subjects[6]

241 assigned to cognitive load to fit separate linear mixed models regressing perceived effort

242 (modeled $n = 1852$) and perceived difficulty ($n = 1848$) on an indicator for whether a subject

243 was recruited on MTurk or from any university. If, as hypothesized, the cognitive load

244 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.

MULTISITE REPLICATION OF TEMPTING FATE   15

245 difficulty might be lower for MTurk subjects. In contrast, perceived effort associated with

246 the cognitive load task was comparable for MTurk and university subjects ($b = 0.63$ with

247 95% CI: [-0.42, 1.68]; $p = 0.24$), as was perceived difficulty ($b = 0.51$ with 95% CI: [-0.11,

248 1.14]; $p = 0.11$). Ultimately, these results do not suggest reduced efficacy of the cognitive

249 load manipulation when implemented online versus in person.

250   To assess differences in academic attitudes, we used subjects[7] from all types of sites,

251 including MTurk, to fit linear mixed models regressing perceived importance ($n = 4175$) and

252 perceived negativity ($n = 4172$) on site type (similar, dissimilar, or MTurk) with random

253 intercepts by site. Contrary to our speculation, MTurk subjects reported, if anything, that

254 answering questions correctly was somewhat more important than did subjects at similar

255 universities ($b = 1.02$ with 95% CI: [0.45, 1.59]; $p = 4.60 \times 10^{-4}$) or at dissimilar universities

256 ($b = 0.76$ with 95% CI: [0.24, 1.29]; $p = 4.60 \times 10^{-3}$). Additionally, when asked to assess how

257 bad it would be to answer incorrectly, MTurk subjects responded comparably to subjects at

258 similar sites ($b = -0.03$ with 95% CI: [-0.52, 0.45]; $p = 0.89$) and at dissimilar sites ($b = 0.45$

259 with 95% CI: [0.01, 0.90]; $p = 0.05$).

260   Lastly, in a planned analysis, we assessed variation in results according to a site's

261 similarity to Cornell, now redefining similarity using a continuous proxy (namely, a

262 university's estimated median total SAT score in 2018) rather than the dichotomous "similar"

263 versus "dissimilar" eligibility criterion for primary analyses. Subjects from universities

264 outside the United States or from MTurk were excluded from this analysis, leaving an

265 analyzed $n = 975$ from 7 universities with median SAT scores ranging from 1182 to 2178 of

266 2400 possible points. We assumed that universities with higher SAT scores would be most

267 similar to Cornell (median SAT: 2134) and therefore considered a linear effect of median

268 SAT score as a moderator of the main effects and interaction of tempting fate with cognitive

269 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.

270 effect of tempting fate ($b = 0.00$ for a 10-point increase in SAT score with 95% CI: [-0.01,

271 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

273   As mentioned in the Introduction, an independent multisite replication of this

274 experiment (Many Labs 2, or "ML2") found strong evidence for a main effect of tempting

275 fate that was in the same direction as that of the original study, but of smaller size; this

276 finding stands in contrast to the present study's negligible main effect estimate. We

277 corresponded with the lead author of ML2 to identify protocol differences that might explain

278 the discrepant results. We identified minor differences in questionnaire design, sampling

279 frame, and statistical analysis (Supplementary Table 4) and pursued post hoc, exploratory

280 analyses to gauge whether these differences were likely to explain the discrepant results (see

281 Supplement for details). For example, we compared results between the two studies' United

282 States MTurk samples to identify questionnaire design effects while holding constant the

283 subject population, we reanalyzed the raw data from both studies in order to eliminate

284 differences in statistical analyses, and we reanalyzed the raw data after redefining the studies'

285 sampling frames to be more directly comparable. However, the two studies' point estimates

286 remained meaningfully different in all of these analyses, so the source of the discrepancy

287 remains unclear. Finally, we estimated a pooled main effect across the primary analysis sites

288 in both studies, which suggested strong evidence for a small main effect of tempting fate

289 (Supplement).

Conclusion

291   We used an updated replication protocol to replicate Risen and Gilovich (2008)'s Study

292 6 in controlled lab settings at universities chosen for their similarity to the original site. We

293 additionally conducted replications on Amazon Mechanical Turk, as in the previous

294 replication, as well as at less similar universities. This replication project has limitations:

295 first, because the number of similar sites was small, we could not reliably assess variation in

MULTISITE REPLICATION OF TEMPTING FATE   17

296 results across these sites. Second, as in all direct replications, our replication was limited to

297 a single operationalization of the tempting-fate effect; our results do not necessarily

298 generalize to other experimental scenarios, for example.

299   Under the updated protocol in similar sites, we estimated a negligible main effect of

300 tempting fate in the absence of cognitive load as well as a negligible focus interaction

301 between tempting fate and cognitive load. Results did not appear to differ between data

302 collected under the updated protocol in similar sites and data collected under the previous

303 replication protocol on Amazon Mechanical Turk, nor did they differ meaningfully in

304 dissimilar universities. Secondary analyses did not support proposed mechanisms of

305 replication failure (namely, reduced effectiveness of the cognitive load manipulation on

306 MTurk or reduced personal salience of the experimental scenario on MTurk). Post hoc

307 analyses did not provide compelling evidence for statistical inconsistency between the

308 original study and replications under the original protocol for the main effect or for the focus

309 interaction. Ultimately, our results fail to support the tempting-fate effect and interaction

310 and also fail to support proposed substantive mechanisms for the replication failure in Open

311 Science Collaboration (2015). However, it is important to note that our negligible main

312 effect estimate stands in contrast to the small main effect identified in Many Labs 2, which

313 and extensive post hoc analyses were unable to identify the source of the discrepancy.

Contributions

315   CRE conceived the Many Labs project. MBM, CRE, and MCF designed this multisite

316 replication study. MBM and DJBP oversaw administration. MBM planned and conducted

317 statistical analyses (with MCF auditing the code) and wrote the manuscript. The remaining

318 authors collected data, audited site-level analyses, and approved the final manuscript. The

319 authors have no conflicts of interest with respect to the authorship or publication of this

320 manuscript. All authors approved the final manuscript with one exception (sadly, SP passed

321 away before the manuscript draft was written).

MULTISITE REPLICATION OF TEMPTING FATE   18

MULTISITE REPLICATION OF TEMPTING FATE 19

332 **References**

333 Epstein, S., Lipson, A., Holstein, C., & Huh, E. (1992). Irrational reactions to negative

334 outcomes: Evidence for two conceptual systems. *Journal of Personality and Social*

335 *Psychology, 62*(2), 328.

336 Klein, R. et a. (2017). Many labs 2: Investigating variation in replicability across

337 sample and setting. *Preprint provided by authors.*

338 Mathur, M., & Frank, M. (2012). Replication of "Why people are reluctant to tempt

339 fate" by Risen & Gilovich. retrieved from https://osf.io/nwua6/.

340 Mathur, M., & VanderWeele, T. (2017). New statistical metrics for multisite

341 replications. *Preprint retrieved from https://osf.io/w89s5/.*

342 Open Science Collaboration. (2015). Estimating the reproducibility of psychological

343 science. *Science, 349*(6251), aac4716.

344 Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can

345 help theorists run behavioral experiments. *Journal of Theoretical Biology, 299*, 172–179.

346 Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of*

347 *Personality and Social Psychology, 95*(2), 293.