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. The authors of the original study expressed concern that the cognitive load manipulation may have been less effective when implemented online and that subjects recruited online may have responded differently to the specific experimental scenario chosen for replication. To address both concerns, 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 target 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

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"). This form of irrational thinking was hypothesized to originate from "System 1" processes that use potentially error-prone heuristics to render fast, effortless judgments. In contrast, alternative "System 2" cognitive processes, which rely on slow, deliberative thinking, are thought to sometimes override System 1's heuristic judgments (e.g., Epstein, Lipson, Holstein, & Huh, 1992). Thus, Risen and Gilovich (2008) additionally hypothesized that System 2 processes may help suppress irrational heuristics regarding tempting fate, and thus that under a cognitive load manipulation designed to preoccupy System 2 resources, the effect of tempting fate on subjects' perceived likelihood of a bad outcome would be magnified. 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 target 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 target interaction effect (estimated effect of tempting 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]; $p = 0.04$).

Mathur and Frank (2012) previously attempted to replicate this study as part of a large-scale replication project (Open Science Collaboration, 2015), finding 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 target interaction ($b = 0.03$ with 95% CI: [-1.14, 1.20]; $p = 0.96$). 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.

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

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

[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

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

We designed the updated protocol in collaboration with the original authors and third-party editor Daniel Simons, resulting in two changes. First, we aimed to recruit universities at which academic pressure to answer questions correctly in class would be comparable to pressure at Cornell University, the site of the original study. Hypothesizing that a university's average SAT score may serve as a proxy for such pressure, we collected primary analysis data on undergraduates at United States universities with estimated median SAT scores in at least the 90th percentile nationally, henceforth termed "similar sites". For comparison, Cornell is in approximately the 95th percentile. Second, rather than collecting data online, we collected data with subjects physically present in controlled settings with minimal distractions and reasonable isolation from other subjects. Acceptable protocols included running each subject alone in a quiet laboratory room or running multiple subjects at the same time in a larger room, but in individual cubicles to minimize social distractions. 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 target effect. For sites whose subjects were not expected to speak fluent English,

until preregistration and data collection were complete.

questionnaire materials were translated and verified through independent back-translation.

Sample sizes in the similar sites were chosen to provide, in aggregate, more than 95% power to detect an interaction effect of the size estimated in the original study. Each site additionally attempted to reach this benchmark internally, though in many cases this was not feasible. The MTurk sample size was also chosen to exceed 95% power to detect the reported effect size. 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.

Results

Descriptive analyses

Four similar university sites (University of Pennsylvania, University of California at Berkeley, University of Virginia, and Stanford University) contributed a total of $n = 754$ analyzed subjects[3] to primary analyses; the MTurk sample contributed $n = 2973$ analyzed subjects[4] to primary analyses. An additional 6 dissimilar university sites contributed $n = 714$ analyzed subjects[5] to secondary analyses. Table 1 displays sample sizes, the number of exclusions, and protocol characteristics for all sites.

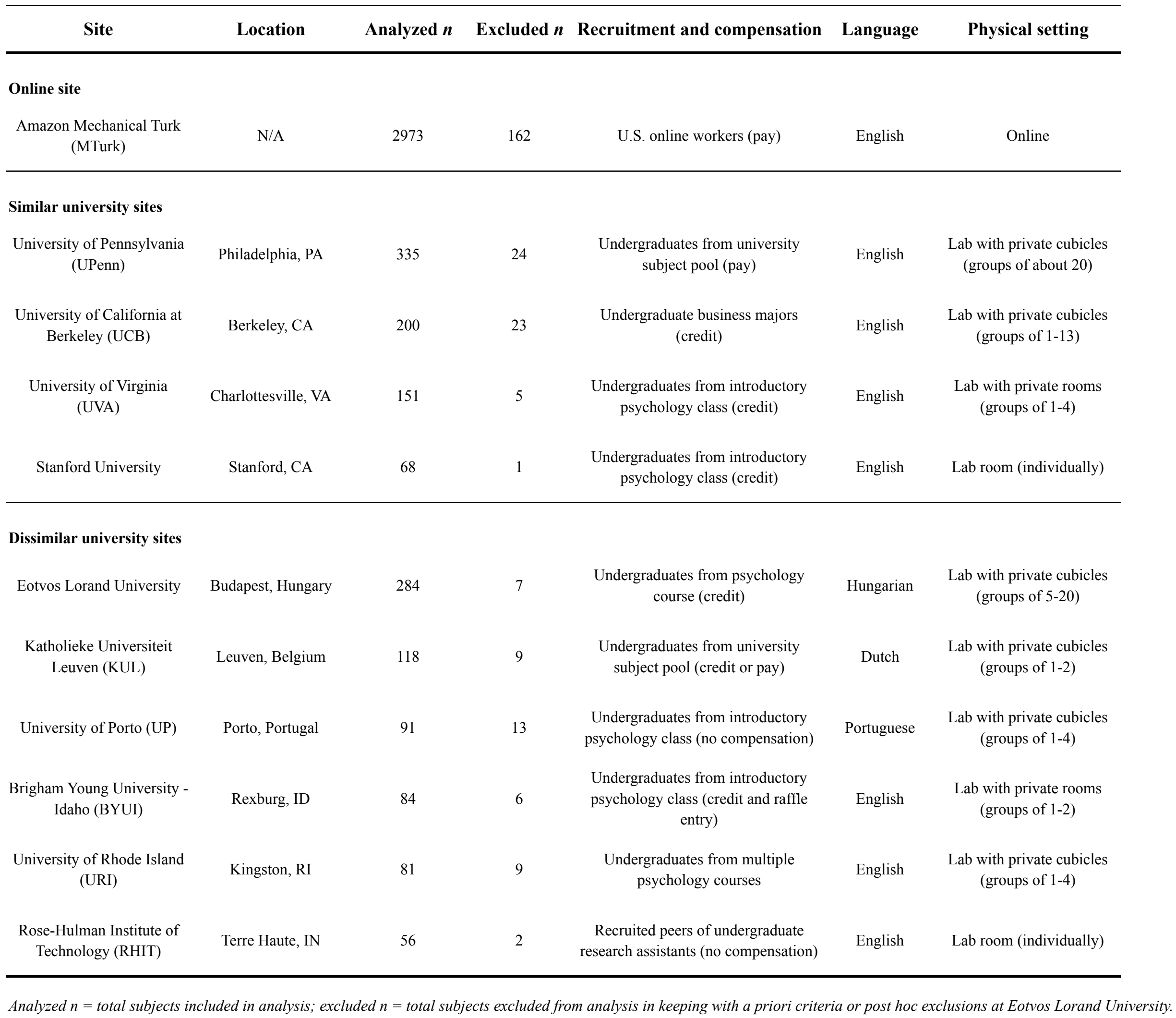
To estimate the main effect of tempting fate and the target interaction within each site, we fit an ordinary least squares regression model of perceived likelihood on tempting fate, cognitive load, and their interaction within each site. This analysis approach is statistically equivalent to the ANOVA 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.[6]

[3] After excluding 7% of the data per *a priori* criteria.

[4] After excluding 5% of the data per *a priori* criteria.

[5] After excluding 6% of the data between *a priori* criteria and an additional, unplanned $n = 7$ exclusions of subjects at Eotvos Lorand University who may have completed the experiment twice.

[6] An alternative for the study-specific estimates would be to use estimates of random intercepts and random



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 1: Summary of sites and participants.

Among the 4 similar sites, 3 had main effect estimates in the same direction as the original study estimate, albeit of considerably smaller magnitude ($b = 0.23$ at University of Pennsylvania, $b = 0.67$ at Stanford, and $b = 0.04$ at University of Virginia vs. 1.03 in the original study). Main effect estimates in similar sites had p-values ranging from 0.30 to 0.94. In the MTurk sample, the target estimate was in the same direction as the original, but was of smaller size, and it was almost identical to the estimate previously obtained under the

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.

same protocol in RPP (0.21 in the present sample vs. 0.20 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 and with confidence intervals substantially overlapping zero with the exception of Eotvos Lorand University, which estimated a main effect comparable to that of the original study ($b = 1.06$ with 95% CI: [0.37, 1.75]; $p = 2.40 \times 10^{-3}$).

Considering the target interaction estimate across sites, only 2 of 4 similar sites had estimates in the same direction as the original, and again, these were of considerably smaller magnitude ($b = 0.17$ at University of Pennsylvania and $b = 0.10$ at University of Virginia vs. 1.54 in the original study). Interaction estimates in similar sites had p-values ranging from 0.43 to 0.89. In the MTurk sample, the target estimate was in the opposite direction from the original estimate and was slightly larger in magnitude than the RPP estimate (-0.20 in the present sample vs. 0.03 in RPP). 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), 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 (b = -0.99 with 95% CI [-1.96, -0.01]; $p = 0.05$).

Primary analyses

Primary analyses aimed to: (1) estimate the target interaction and the main effect under the updated protocol in similar sites; and (2) assess whether the target 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 (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

Table 2: In units of perceived likelihood on a 0-10 scale, estimates of the main effect and target 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

random effects were assumed independently and identically normal.[7] This model allows estimation of the target 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 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 2), and nor did results collected under the updated protocol in similar sites (Table 2, row 2). Updating the protocol did not appear to

[7] 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).

change the main effect estimate (Table 2, row 3). Furthermore, results from the new MTurk sample also did not support the target interaction (Table 2, row 4), and nor did results under the updated protocol (Table 2, row 5). As seen for the main effect, updating the protocol did not meaningfully affect the target interaction estimate (Table 2, row 6). Both the main effect of tempting fate and the target 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 incorporating data from dissimilar university sites (total $n = 4441$). Site type was treated as a categorical variable (MTurk, similar university site, or dissimilar university site)[8]. Additionally, these analyses formally estimated the difference in results between similar and dissimilar sites. Results (Table 3) did not support the main effect or the target interaction in any site type. The main effect estimate in dissimilar sites was comparable to that in similar sites (Table 3, row 4), as was the interaction estimate (Table 3, 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).

Statistical consistency of replication with original results

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 & 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

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

Table 3: In units of perceived likelihood on a 0-10 scale, estimates of the main effect and target 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

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 target 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.

Evaluating proposed explanations for replication failure

Anticipating that results may have differed between similar and dissimilar sites, we had planned to conduct secondary analyses assessing evidence for whether these differences were attributable to the original authors' hypotheses regarding 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, 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 target 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]; $p = 0.95$).

We also collected two new measures, developed through discussion with the original authors, in which 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?"*). These provided manipulation checks of whether the cognitive load manipulation was effortful and difficult, as intended. We used subjects[9] 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

[9] 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.

settings, perceived effort or 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.

The original authors also 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. Thus, the latter subjects may respond differently. 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?"*). We used subjects[10] 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]; $p = 0.05$).

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

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 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 target interaction ($b = 0.00$ with 95% CI: [-0.02, 0.02]; $p = 0.97$).

Conclusion

We used an updated replication protocol, developed in collaboration with the original authors, 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. Under the updated protocol in similar sites, we estimated a negligible main effect of tempting fate in the absence of cognitive load (regression coefficient estimate $b = 0.11$ with 95% CI: [-0.34, 0.56]; $p = 0.64$ vs. in the original study: $b = 1.03$ with 95% CI: [0.09, 1.97]; $p = 0.03$) as well as a negligible target interaction between tempting fate and cognitive load ($b =$ -0.02 with 95% CI: [-0.68, 0.63]; $p = 0.94$ vs. in the original study: $b =$ 1.54 with 95% CI: [0.05, 3.03]; $p = 0.04$). 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 ($P_{orig} = 0.12$) or for the target interaction ($P_{orig} = 0.07$). Limited power in the original study likely accounts for the apparent discrepancy between, on the one hand, negligible effect sizes estimated in the replications and, on the other hand, fairly modest statistical inconsistency between the original study and the replications. Ultimately, our results fail to support the tempting-fate effect and interaction and also fail to support proposed substantive mechanisms for the replication failure.

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.

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.

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.

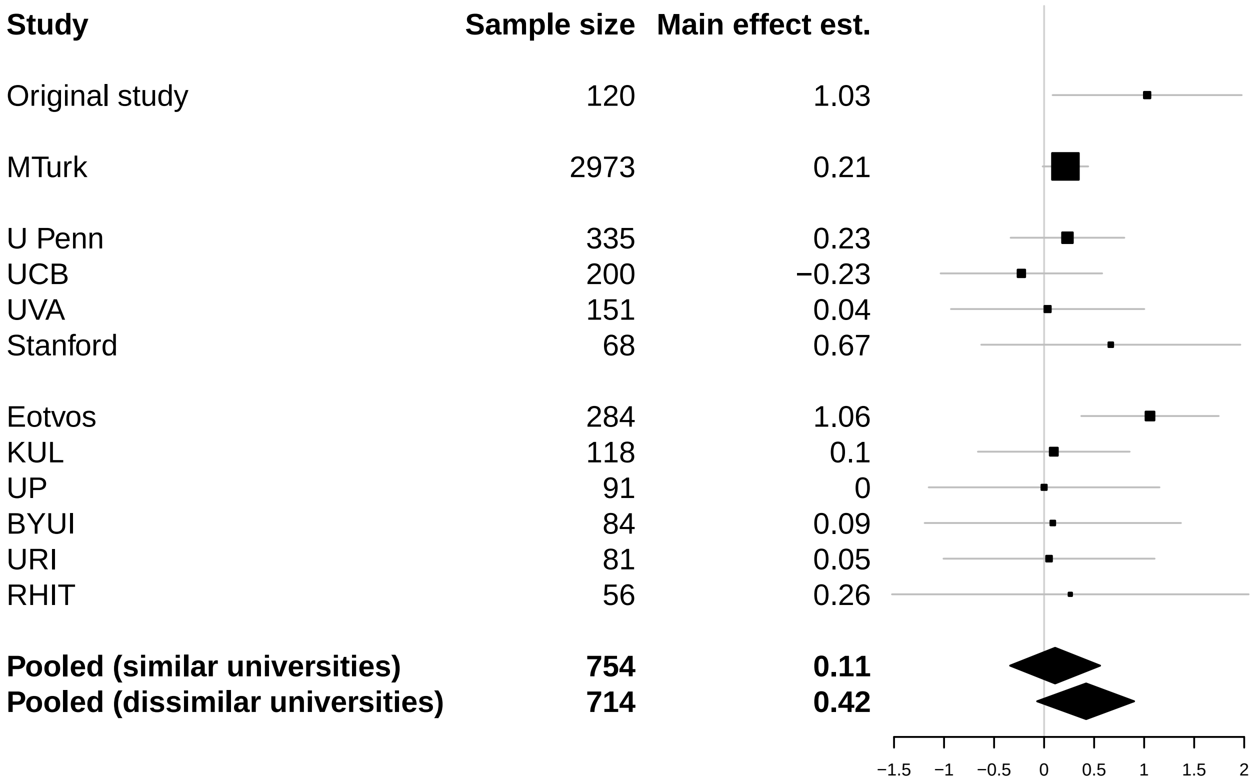
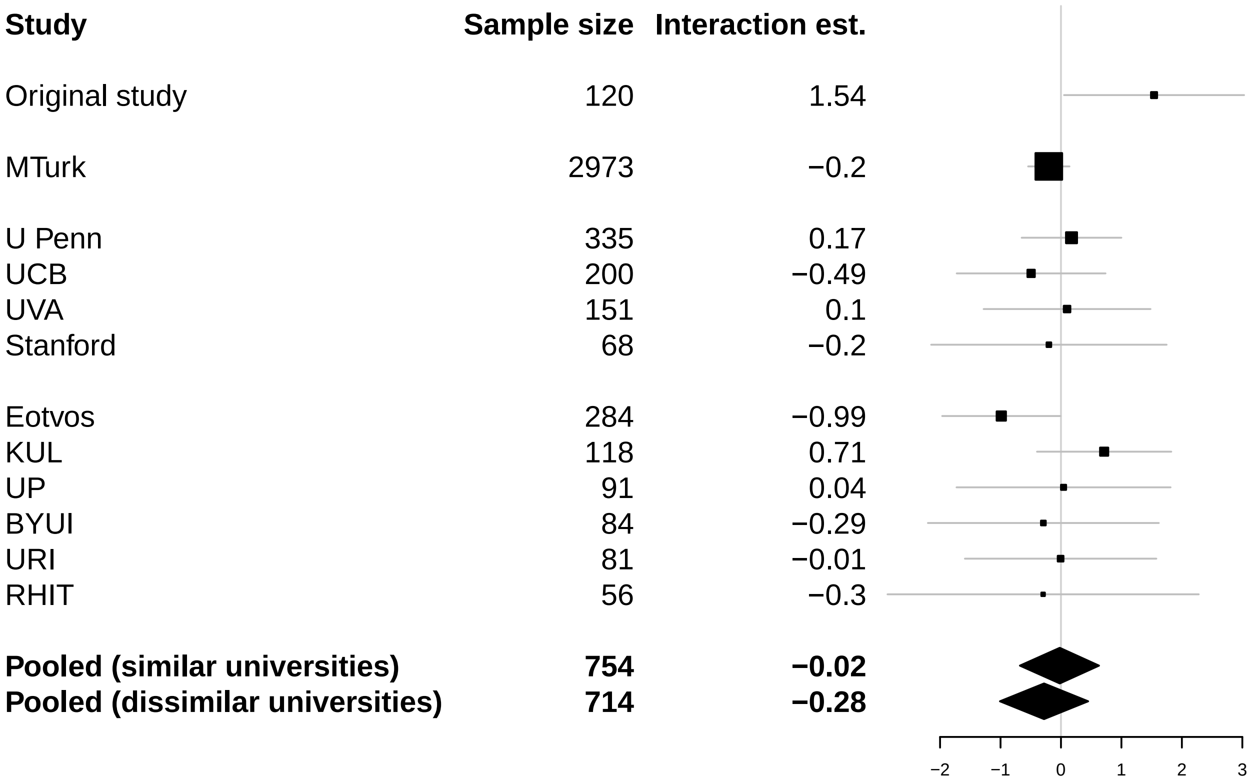


Figure 1: Forest plot for main effect estimates ordered by site type (MTurk, similar, dissimilar) and then by sample size. Point estimates and 95% CIs for each site are from ordinary least squares regression fit to that site's data. For similar sites, pooled point estimates and 95% CIs for similar sites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



Study	Sample size	Interaction est.
Original study	120	1.54
MTurk	2973	−0.2
U Penn	335	0.17
UCB	200	−0.49
UVA	151	0.1
Stanford	68	−0.2
Eotvos	284	−0.99
KUL	118	0.71
UP	91	0.04
BYUI	84	−0.29
URI	81	−0.01
RHIT	56	−0.3
Pooled (similar universities)	**754**	**−0.02**
Pooled (dissimilar universities)	**714**	**−0.28**

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 are from ordinary least squares regression fit to that site's data. For similar sites, pooled point estimates and 95% CIs for similar sites are from the primary mixed model. For dissimilar sites, these are from the secondary mixed model. Pooled point estimates represent the average interaction effect among subjects in similar universities or in all universities.