Failure to replicate tempting-fate effects:

registered multisite replication of Risen & Gilovich (2008)

Maya B. Mathur[1,2*] *and FRIENDS*[1,3]

[1] Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA

[2]Quantitative Sciences Unit, Stanford University, Palo Alto, CA, USA

[3]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA

∗: Corresponding author:

mmathur@stanford.edu

Quantitative Sciences Unit (c/o Inna Sayfer)

1070 Arastradero Road

Palo Alto, CA

94305

Notes to self

- When submitting: Separate appendix and main text; remember to submit Table 1 as separate Excel file!
- Ask Charlie for funding statement
- Fill in "XXX-point scale" thing
- Why are forest plots cut off at the bottom??
- Put sample sizes for all models and cross-check against df
- Add authors: Diane-Jo is working on this

Abstract

WC: 249/250

Risen et al. (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 using Amazon Mechanical Turk. The original study's authors 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. To address both concerns, we developed a new protocol in collaboration with the original authors. We used four university sites 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 existence of the target interaction or the main effect and changed very little when including an additional six universities that were less similar to the original site. Post hoc analyses were weakly suggestive of statistical inconsistency between the original study's estimates and the replications; that is, the original study's results would have been fairly unlikely in the estimated distribution of the replications. We also collected a new Mechanical Turk sample under the previous replication protocol to determine 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. Planned secondary analyses failed to support substantive mechanisms for the failure to replicate.

Introduction

Risen et al. (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 et al. (1992)). Thus, Risen et al. (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 statistical interaction between cognitive load and tempting fate on subjects' perceived likelihood of an ironic bad outcome.

Risen et al. (2008)'s Study 6, the target of replication, used a between-subjects factorial design to assess this possibility by manipulating the behavior of a character in a scenario (a student who had either tempted fate by not doing the assigned reading or who had not tempted fate) as well as the presence or absence of cognitive load on subjects. Subjects assigned to complete the task without cognitive load simply read the scenario and then judged the likelihood of being called on in class. Subjects assigned to complete the task under cognitive load were required to count 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 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 effort (Open Science Collaboration, 2015), finding little evidence for either a main effect of tempting fate without cognitive load[2] ($b = 0.20$ with 95% CI: [-0.58, 0.97]; $p = 0.62$) or the target interaction ($b = 1.54$ with 95% CI: [-1.14, 1.20]; $p = 0.96$). However, prior to the collection of replication data, the authors of the original study expressed concerns about the replication protocol. Specifically, the replication was implemented on the crowdsourcing website Amazon Mechanical Turk, a setting that could 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

[1] Approximate effect sizes were recomputed from rounded values in (Risen et al., 2008).

[2] For both the original study and the replications, we describe inference on the main effect of tempting fate under the "Type III" sum-of-squares decomposition rather than the "Type I" decomposition sometimes used in F-tests. The latter is not invariant to the order in which the covariates are included in the model and lacks a straightforward interpretation in terms of differences of subgroup means.

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, the present multisite replication project aimed to: (1) reassess replicability of (Risen et al., 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.

Methods

The protocol, sample size criteria, exclusion criteria, and statistical analysis plan were preregistered[3] with details publicly available (https://osf.io/h5a9y/); any departures from these plans are reported in this manuscript. We designed the updated protocol in collaboration with the original authors and editor Daniel Simons, resulting in the following changes. First, to more closely approximate the sampling frame of the original study, which was conducted on Cornell University undergraduates, we collected our 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 replication protocol (Open Science Collaboration, 2015) ("RPP") without modification to collect a new sample on Amazon Mechanical Turk ("MTurk"). Finally, we collected secondary data in several universities located outside the United States or not meeting the SAT criterion for similarity to Cornell, 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, questionnaire materials were translated and verified through independent back-translation.

Sample sizes in the similar sites were chosen to allow, 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 performed by one author (MM), who was blinded to results until all sites had completed data collection; these analyses were audited for accuracy by other authors.

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

Descriptive results

Four similar university sites (University of Pennsylvania, University of California at Berkeley, University of Virginia, and Stanford University) contributed a total of $n = 763$ analyzed subjects (after exclusions per *a priori* criteria) to primary analyses; the MTurk sample contributed $n = 2973$ analyzed subjects to primary analyses. An additional 6 dissimilar university sites contributed $n = 714$ analyzed subjects 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, respectively[4]. 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 $b = 0.03$ at University of Virginia vs. 1.03 in the original study). Main effect estimates in similar sites had p-values ranging from 0.31 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 same protocol by 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 obtained a main effect comparable to that of the original study ($b = 1.06$ with 95% CI: 0.37, 1.75; $p = 0.003$).

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 $b = 0.09$ 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 previous estimate obtained under the same protocol (-0.2 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; $p = 0.05$).

[4] 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.

Replication results under the updated protocol

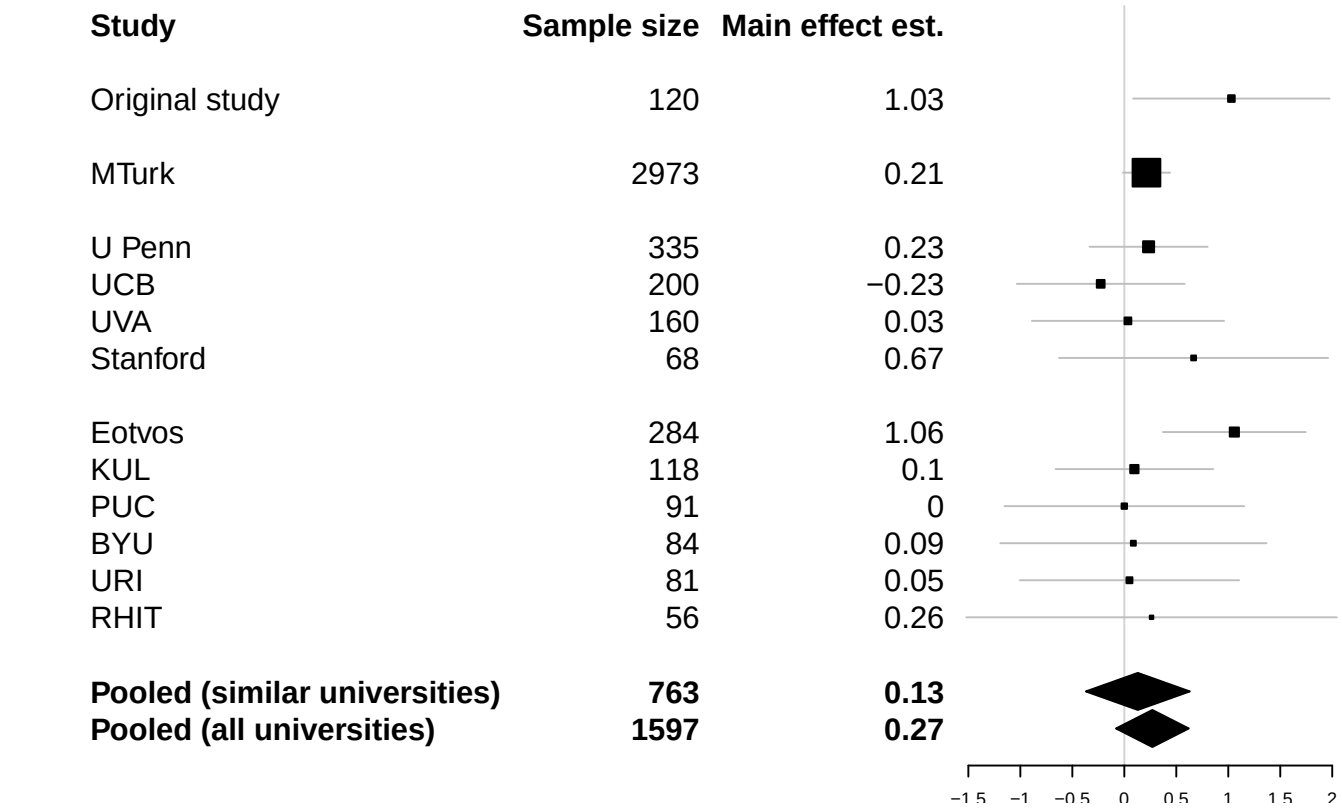


Figure 1: Forest plot for main effect estimates ordered by site type (MTurk, similar, dissimilar) and then by sample size. Point estimates and 95% CIs for each site are from ordinary least squares regression fit to that site's data. Point estimates and 95% CIs for pooled estimates are from primary and secondary mixed models.

Primary analyses aimed to: (1) estimate the target interaction and the main effect under the updated protocol in similar sites; and (2) asssess 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 random effects were assumed independently and identically normal[5]. 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 (https://osf.io/h5a9y/)).

[5]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 (Appendix).

Study	Sample size	Main effect est.
Original study	120	1.03
MTurk	2973	0.21
U Penn	335	0.23
UCB	200	−0.23
UVA	160	0.03
Stanford	68	0.67
Eotvos	284	1.06
KUL	118	0.1
PUC	91	0
BYU	84	0.09
URI	81	0.05
RHIT	56	0.26
Pooled (similar universities)	**763**	**0.13**
Pooled (all universities)	**1597**	**0.27**

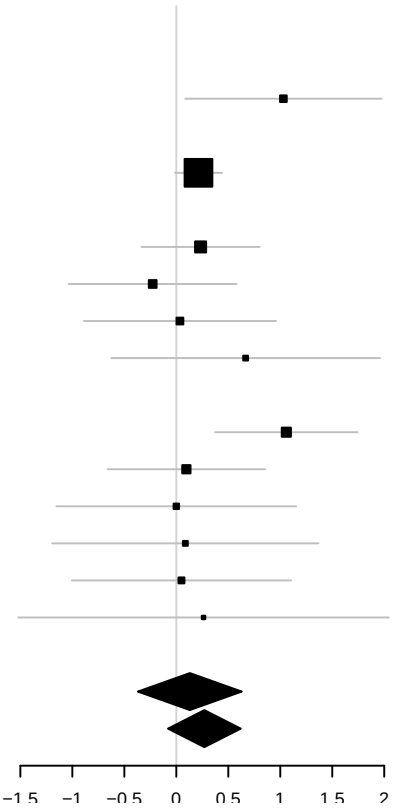


Figure 2: Forest plot for main effect estimates ordered by site type (MTurk, similar, dissimilar) and then by sample size. Point estimates and 95% CIs for each site are from ordinary least squares regression fit to that site's data. Point estimates and 95% CIs for pooled estimates are from primary and secondary mixed models.

Parameter	Estimate	95% CI	p-value
Magnitude of X main effect within MTurk	0.21	[-0.27, 0.70]	0.67
Magnitude of X main effect within similar sites	0.13	[-0.37, 0.63]	0.62
Effect of similar site vs. MTurk on X main effect	-0.09	[-0.78, 0.61]	0.85
Magnitude of X-L interaction within MTurk	-0.20	[-0.76, 0.35]	0.73
Magnitude of X-L interaction within similar sites	-0.03	[-0.72, 0.67]	0.94
Effect of similar site vs. MTurk on X-L interaction	0.17	[-0.72, 1.06]	0.75

Table 2: Estimates of the main effect and target interaction effect under the updated protocol (similar sites) and the RPP protocol (MTurk), as well as estimates of the difference between these estimates. X = tempting fate; L = cognitive load.

Consistent with the RPP replication, the present results collected under the updated protocol in similar sites did not support the main effect of tempting fate (regression coefficient estimate $b = 0.13$ with 95% CI: [-0.37, 0.63]; $p = 0.62$), and nor did results from MTurk ($b = 0.21$ with 95% CI: [-0.27, 0.70]; $p = 0.67$). Updating the protocol did not appear to change the main effect estimate ($b = -0.09$ with 95% CI: [-0.78, 0.61]; $p = 0.85$). Furthermore, results under the updated protocol also did not support the target interaction (regression coefficient estimate $b = -0.03$ with 95% CI: [-0.72, 0.67]; $p = 0.94$), nor did results from the new MTurk sample collected under the RPP protocol ($b = -0.20$ with 95% CI: [-0.76, 0.35]; $p = 0.73$). As for the main effect, updating the protocol did not meaningfully affect the target interaction estimate ($b = 0.17$ with 95% CI: [-0.72, 1.06]; $p = 0.75$). Both the main effect of tempting fate and the target interaction appeared relatively homogeneous across sites (estimated random intercept standard deviation = 0.22; estimated random slope standard deviation = 0.23).

Replication results in all university sites

A planned secondary analysis addressed the same questions as the primary analyses, but using data from all university sites rather than only similar sites. Results (Table 3) were qualitatively similar to primary results, again providing little support for the main effect of tempting fate, the target interaction, or differences between the estimates in university sites versus under the RPP protocol. As in primary analyses, the main effect and interaction appeared relatively homogeneous across sites (estimated random intercept standard deviation = 0.23; estimated random slope standard deviation = 0.38).

Parameter	Estimate	95% CI	p-value
Magnitude of X main effect within MTurk	0.21	[-0.28, 0.71]	0.40
Magnitude of X main effect within university sites	0.27	[-0.08, 0.62]	0.15
Effect of university site vs. MTurk on X main effect	0.06	[0.67, -0.55]	0.87
Magnitude of X-L interaction within MTurk	-0.20	[-1.01, 0.60]	0.62
Magnitude of X-L interaction within university sites	-0.17	[-0.69, 0.35]	0.53
Effect of university site vs. MTurk on X-L interaction	0.03	[0.99, -0.93]	0.95

Table 3: Estimates of the main effect and target interaction effect in all university sites and under the RPP protocol (MTurk), as well as estimates of the difference between these estimates. X = tempting fate; L = cognitive load.

Statistical consistency of replication results with original results

```
##
## The p-value of the original study under the null hypothesis of original-replication consistency is:

##
## The p-value of the original study under the null hypothesis of original-replication consistency is:

##
## The p-value of the original study under the null hypothesis of original-replication consistency is:

##
## The p-value of the original study under the null hypothesis of original-replication consistency is:
```

To supplement primary analyses, which focused on using the replication data to re-estimate the target effect size, 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 (Mathur and VanderWeele, 2017). 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, 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 ($P_{orig} = 0.18$) 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 $P_{orig} = 0.08$; this probability decreases slightly to $P_{orig} = 0.04$ when considering the distribution of all university sites.

Evaluating proposed explanations for the replication failure

In planned secondary analyses, we assessed the original authors' hypotheses regarding the previous replication failure in RPP. 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 was recruited on MTurk or from any university (details in Appendix). 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 (-0.03 with 95% CI: -0.99, 0.93; $p = 0.95$).

TEST: 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 XXX-point scale the perceived effort associated with this task (*"How much effort did the counting task require?"*) and its 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[6] assigned to cognitive load to fit separate linear mixed models regressing perceived effort ($n = 1857$) and perceived difficulty ($n = 1853$) 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 difficulty might be lower for MTurk subjects. In contrast, perceived effort of the cognitive load task was comparable for MTurk vs. university subjects ($b = 0.62$ with 95% CI: -0.43, 1.67; $p = 0.30$), as was perceived difficulty ($b = 0.50$ with 95% CI: -0.12, 1.13; $p = 0.24$). Ultimately, these analyses do not suggest reduced efficacy of the cognitive load manipulation when implemented online versus in person.

[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.

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 subjects which required subjects to evaluate the importance of answering questions correctly in class (*"If you were a student in the scenario you just read about, how important would it be for you to answer questions correctly in class?"*) and 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[7] from all types of sites, including MTurk, to fit linear mixed models regressing perceived importance ($n = 4184$) and perceived negativity ($n = 4181$) on site type (similar, dissimilar, or MTurk) with random intercepts by site. Contrary to prediction, 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, 1.58]; $p = 0.07$) or at dissimilar universities ($b = 0.76$ with 95% CI: [0.24, 1.28]; $p = 0.10$). Additionally, when asked to assess how bad it would be to answer incorrectly, MTurk subjects responded comparably to subjects at similar sites ($b = -0.01$ with 95% CI: [-0.51, 0.49]; $p = 0.98$) and at dissimilar sites ($b = 0.46$ with 95% CI: [0.00, 0.91]; $p = 0.25$).

Lastly, 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 $n = 984$. We assumed that universities with higher SAT scores would be most similar to Cornell (median SAT: 2134 of 2400 possible) 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; $p = 0.84$) 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 original authors to replicate Risen et al. (2008)'s Study 6 in controlled lab settings at universities chosen for their similarity to the original site. We additionally ran replications on Amazon Mechanical Turk, as in the previous replication, and in less similar universities. Under the updated protocol in similar sites, we estimate only a negligible main effect of tempting fate (regression coefficient estimate $b = 0.13$ with 95% CI: [-0.37, 0.63]; $p = 0.62$ vs. in the original study: $b = 1.03$ with 95% CI: [0.09, 1.97];

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

$p = 0.03$) as well as a negligible target interaction between tempting fate and cognitive load (b = -0.03 with 95% CI: [-0.72, 0.67]; $p = 0.94$ vs. in the original study: $b = 1.54$ with 95% CI: [0.05, 3.03]; $p = 0.04$). We found no evidence to suggest that estimates of the main effect differed between data collected under the updated protocol in similar sites and data collected under the previous replication protocol on Amazon Mechanical Turk, nor did results change meaningfully when including less similar universities in analysis. Planned secondary analyses did not support proposed mechanicams 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 suggested weak evidence of statistical inconsistency between the original study and replications under the original protocol for the main effect ($P_{orig} = 0.12$) and for the target interaction ($P_{orig} = 0.08$). Ultimately, our results fail to replicate the original study and additionally fail to support proposed substantive mechanisms for the replication failure.

Funding

Acknowledgments

Supplementary Analysis Code

Contents

Data Quality

Excluded and analyzed subjects by site

site	n.excl	n.total
MTurk	162	2,973
U Penn	24	335
UCB	23	200
UVA	5	160
Stanford	1	68
Eotvos	7	284
KUL	9	118
PUC	13	91
BYU	6	84
URI	9	81
RHIT	2	56

```
# sample sizes by site type
t = table( b$group )
stargazer( as.data.frame(t), header=FALSE, summary=FALSE,
           rownames = FALSE,
           colnames = FALSE,
           title = "Total analysis sample sizes by site type" )
```

Total analysis sample sizes by site type

a.mturk	2,973
b.similar	763
c.dissimilar	714

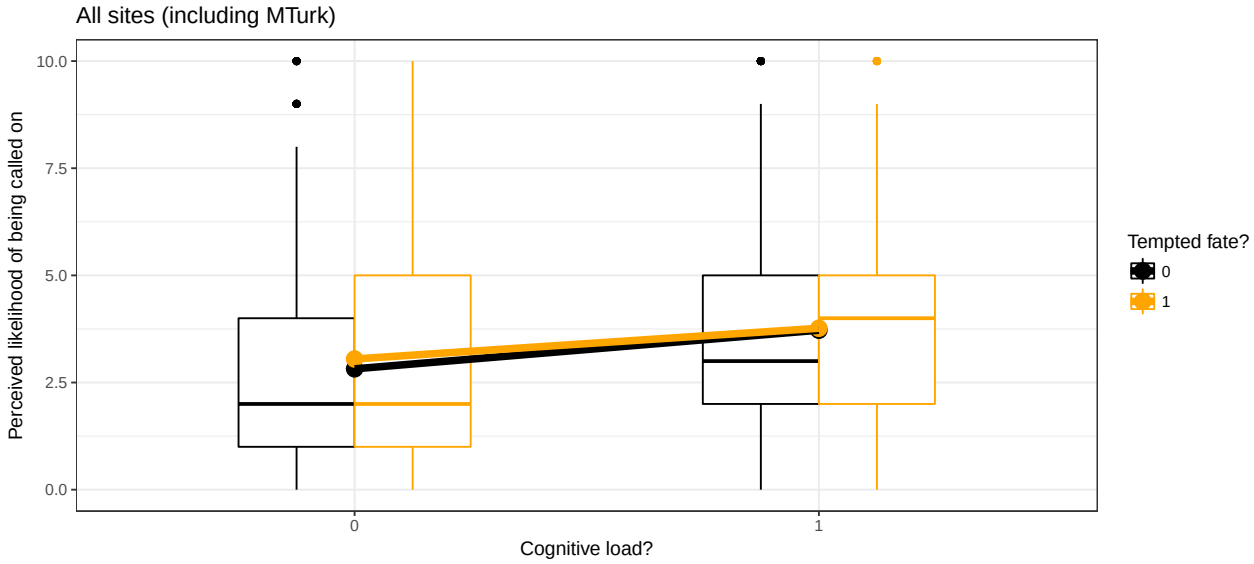
We excluded subjects exactly per the preregistration and the original study protocol, resulting in 261 exclusions across all sites, including MTurk. This is 6% of the originally collected data.

Descriptive Stats and Plots

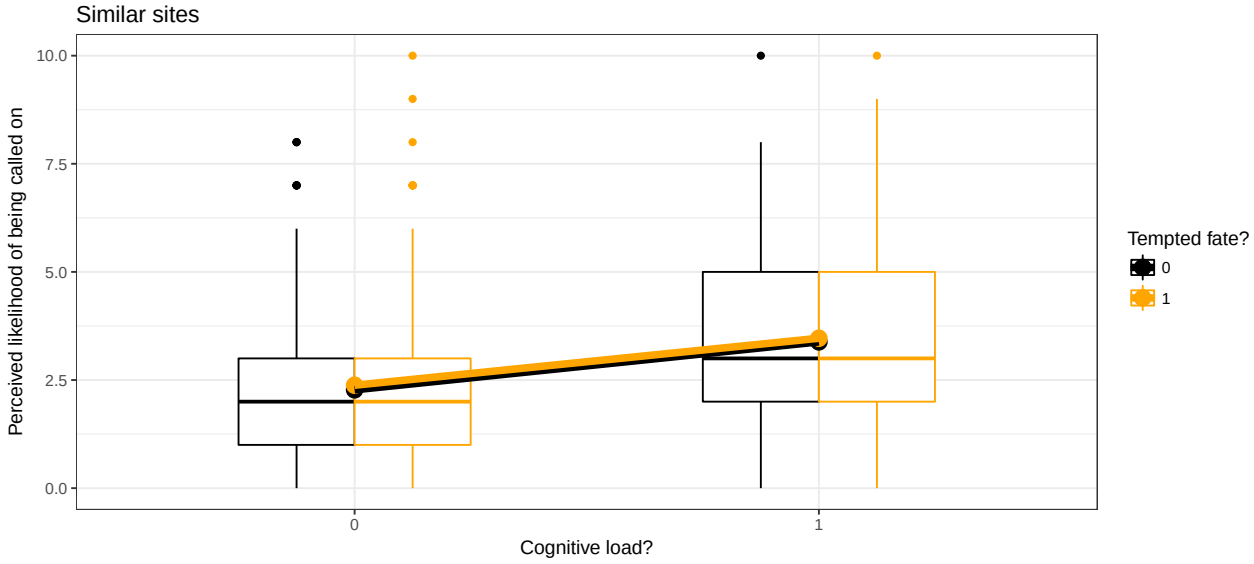
Boxplots: medians and IQRs; lines: simple means by subset. (For the same plots within each site, see the data prep PDF.) These aggregated means and SDs pool across all sites within a group (similar, dissimilar, MTurk) and do not account for clustering by site.

```
##### Fn: Interaction Plot #####
# pass the desired subset of data
int_plot = function( dat, ggtitle ) {
    agg = ddply( dat, .(load, tempt), summarize, val = mean(lkl, na.rm=TRUE)   )  # aggregate data for plotting hap

  colors = c("black", "orange")
  plot( ggplot( dat, aes(x = as.factor(load), y = lkl, color=as.factor(tempt) ) ) + geom_boxplot(width=0.5) +
      geom_point(data = agg, aes(y = val), size=4 ) +
      geom_line(data = agg, aes(y = val, group = tempt), lwd=2 ) +
      scale_color_manual(values=colors) +
      scale_y_continuous( limits=c(0,10) ) +
      ggtitle(ggtitle) +
```

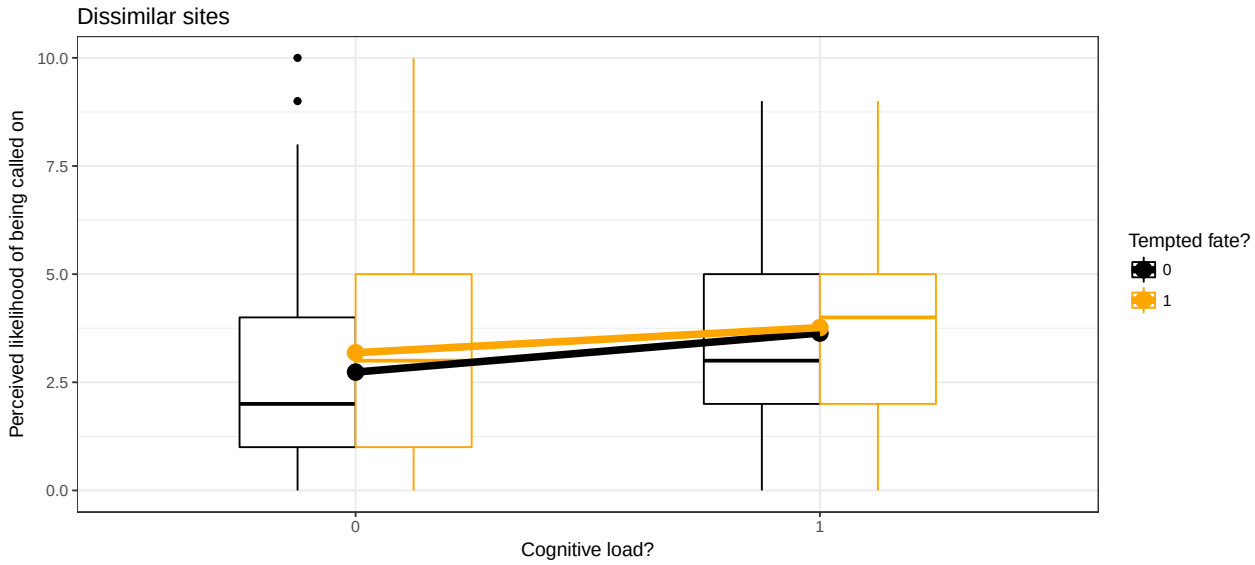
```
      theme_bw()  + xlab("Cognitive load?") + ylab("Perceived likelihood of being called on") +
      guides(color=guide_legend(title="Tempted fate?"))
    )
}

##### Plot By Subset #####
int_plot(b, ggtitle = "All sites (including MTurk)")  # all sites
```



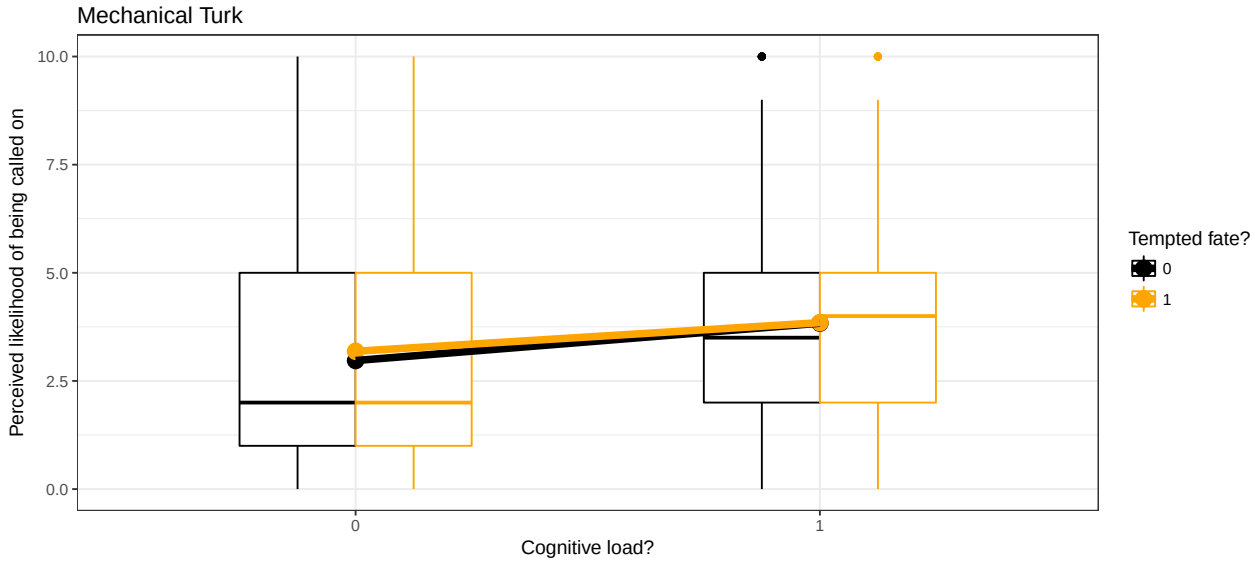

```
int_plot(b[ b$group=="b.similar", ], ggtitle = "Similar sites")
```




```
int_plot(b[ b$group=="c.dissimilar", ], ggtitle = "Dissimilar sites")
```




```
int_plot(b[ b$group=="a.mturk", ], ggtitle = "Mechanical Turk")
```



Means and SDs by site type

```
agg.means = aggregate( lkl ~ tempt + load + group, b, mean)
agg.sds = aggregate( lkl ~ tempt + load + group, b, sd)

agg = data.frame( cbind( agg.means, agg.sds$lkl) )
names(agg)[4:5] = c("mean", "SD")

stargazer( agg, header=FALSE, summary=FALSE,
           title = "Means and SDs of perceived likelihood across all subjects
           within each site type (naively pooling all sites)" )
```

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

	tempt	load	group	mean	SD
1	0	0	a.mturk	2.972	2.392
2	1	0	a.mturk	3.185	2.410
3	0	1	a.mturk	3.835	2.313
4	1	1	a.mturk	3.847	2.317
5	0	0	b.similar	2.274	1.884
6	1	0	b.similar	2.380	1.957
7	0	1	b.similar	3.382	2.158
8	1	1	b.similar	3.466	2.180
9	0	0	c.dissimilar	2.735	2.020
10	1	0	c.dissimilar	3.184	2.195
11	0	1	c.dissimilar	3.639	2.104
12	1	1	c.dissimilar	3.763	2.007

Sanity Checks 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, try fitting 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 0.13 with 95% CI: (-0.37, 0.63), whereas in the present subset model, it is 0.13 with 95% CI: (-0.33, 0.59). Also, in the primary model, the estimated interaction effect was -0.03 with 95% CI: (-0.72, 0.67), 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, fit random-effects meta-analysis to the point estimates. For the main effect:

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

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

```
##
## The p-value of the original study under the null hypothesis of original-replication consistency is:

## [1] 0.0785936
```

In the mixed model (Model 1), the estimated main effect and heterogeneity in similar sites was $\widehat{M} = 0.13$ and $\widehat{V} = 0.06$ in the mixed model compared to $\widehat{M} = 0.13$ and $\widehat{V} = 0$ in the meta-analysis. They agree very closely. P_{orig} is a bit lower due to the lower estimated heterogeneity here.

For the target interaction effect:

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

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

```
##
## The p-value of the original study under the null hypothesis of original-replication consistency is:

## [1] 0.05280074
```

In the mixed model (Model 1'), the estimated interaction effect and heterogeneity in similar sites was $\widehat{M}$ = -0.03 and $\widehat{V}$ = 0.14 in the mixed model compared to $\widehat{M}$ = -0.02 and $\widehat{V}$ = 0 in the meta-analysis. They agree very closely.

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.

```
summary( aov( lkl ~ load * tempt, data = b[ b$group == "b.similar", ] ) )
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## load          1  228.0  227.98  54.781 3.58e-13 ***
## tempt         1    1.8    1.75   0.421    0.517
## load:tempt    1    0.0    0.02   0.006    0.940
## Residuals   759 3158.8    4.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These results are qualitatively similar to what we saw in the primary model.

Other Planned Models

The above analyses did not suggest differences in results between similar and dissimilar sites. Therefore, as planned in the preregistered protocol, we did not pursue the secondary mediation models.

References

Seymour Epstein, Abigail Lipson, Carolyn Holstein, and Eileen Huh. Irrational reactions to negative outcomes: evidence for two conceptual systems. *Journal of Personality and Social Psychology*, 62(2):328, 1992.

MB Mathur and MC Frank. Replication of "Why people are reluctant to tempt fate" by Risen & Gilovich. 2012. Retrieved from https://osf.io/nwua6/.

MB Mathur and TJ VanderWeele. New statistical metrics for multisite replications. 2017. Preprint retrieved from https://osf.io/w89s5/.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

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

Jane L Risen, Thomas Gilovich, et al. Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology*, 95(2):293, 2008.