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

Contents

Descriptive Statistics and Plots

Standardized mean differences and t-tests within each site

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

Table S1: Standardized mean difference (SMD) and p-values for t-tests of the effect of tempting fate on perceived likelihood, stratified by cognitive load within each site.

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

Interaction plots by site type

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

Figure S1: All sites (including MTurk)

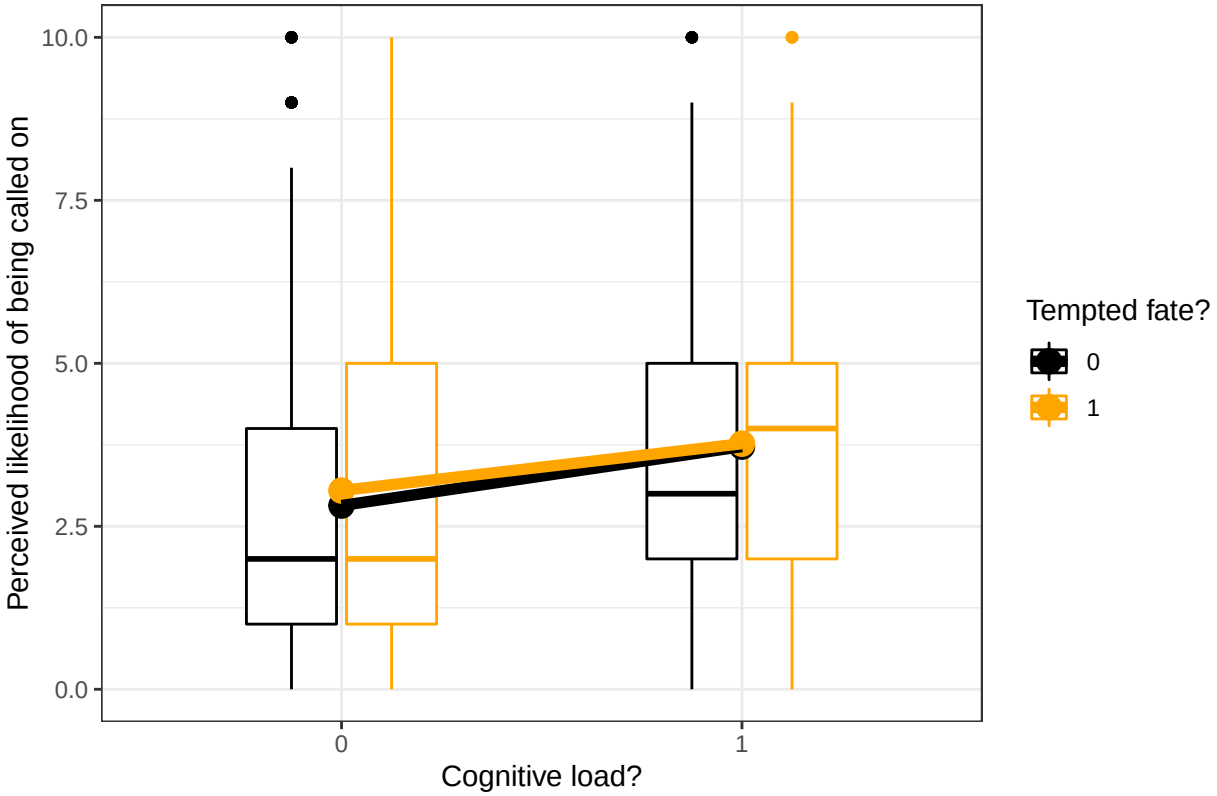


Figure S2: Similar sites

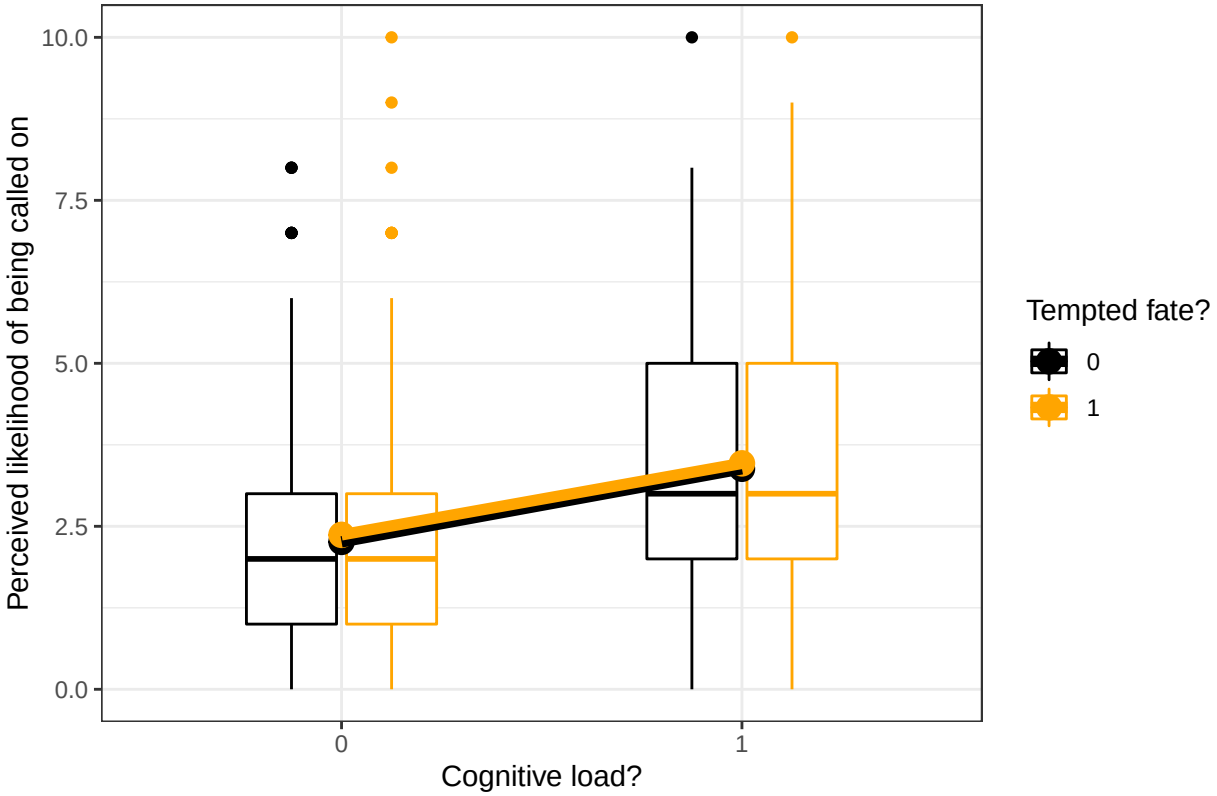


Figure S3: Dissimilar sites

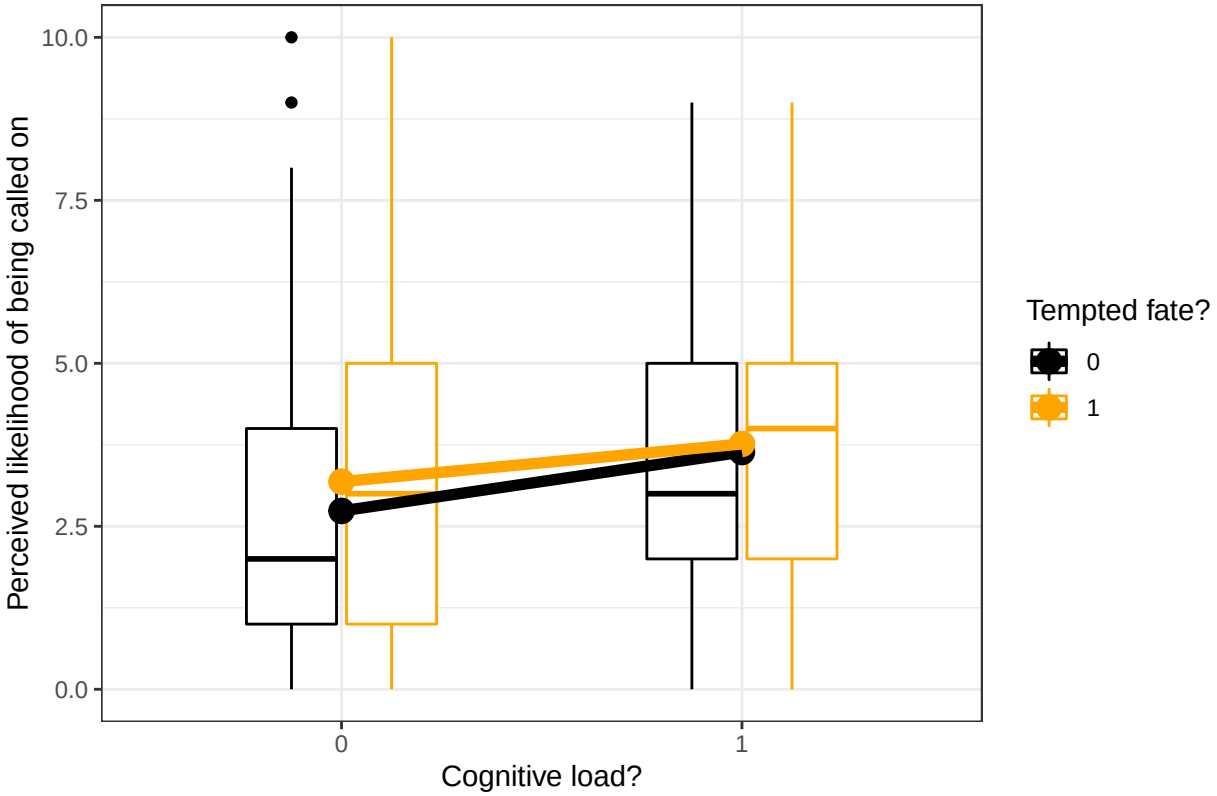
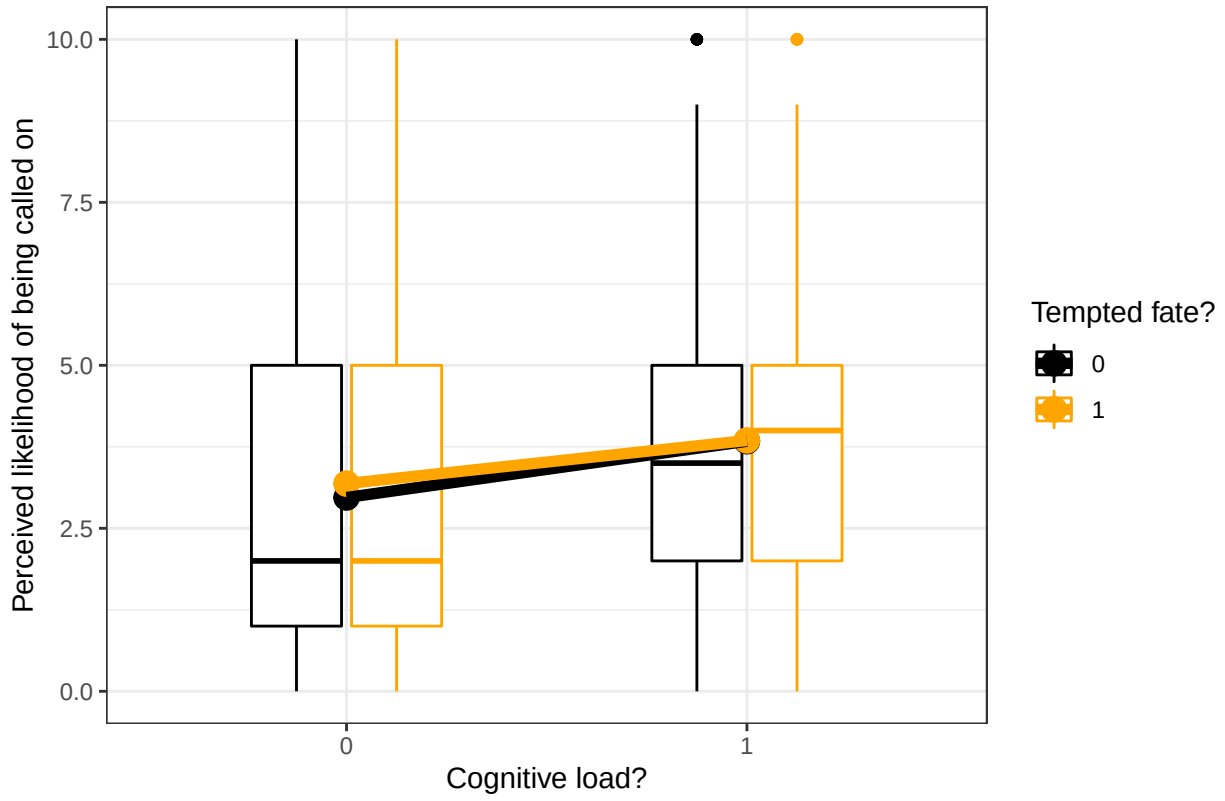


Figure S4: Mechanical Turk



Cell means and standard deviations by site type

Table S2: Means and SDs of perceived likelihood across all subjects within each site type (naively pooling all sites)

Tempt fate	Cognitive load	Group	Mean	SD
0	0	Dissimilar	2.74	2.02
1	0	Dissimilar	3.18	2.19
0	1	Dissimilar	3.64	2.10
1	1	Dissimilar	3.76	2.01
0	0	MTurk	2.97	2.39
1	0	MTurk	3.19	2.41
0	1	MTurk	3.84	2.31
1	1	MTurk	3.85	2.32
0	0	Similar	2.26	1.89
1	0	Similar	2.37	1.96
0	1	Similar	3.38	2.16
1	1	Similar	3.47	2.19

Statistical Consistency of Original Study with Replications

We conducted post hoc secondary analyses to assess the extent to which the replication findings were statistically consistent with the original study; that is, whether it is plausible that the original study was

drawn from the same distribution as the replications (Mathur and VanderWeele 2017). These analyses account for uncertainty in both the original study and the replication and for possible heterogeneity in the replications, and they can help distinguish whether an estimated effect size in the replications that appears to disagree with the original estimate may nevertheless be statistically consistent with the original study due, for example, to low power in the original study or in the replications or to heterogeneity.

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

Sensitivity Analyses for Reported Results

Fit subset model counterpart to primary analysis model

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

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

```
## boundary (singular) fit: see ?isSingular
```

```
## Warning: Model failed to converge with 1 negative eigenvalue: -2.5e+00
```

```
CI.temp = confint( m1.temp, method = "Wald" )
```

In the primary model, the estimated main effect was 0.11 with 95% CI: (-0.34, 0.56), whereas in the present subset model, it is 0.13 with 95% CI: (-0.33, 0.59). Also, in the primary model, the estimated interaction effect was -0.02 with 95% CI: (-0.68, 0.63), whereas in the present subset model, it is -0.02 with 95% CI: (-0.66, 0.61). These results are similar.

Fit meta-analytic counterparts to primary analysis model

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

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

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

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

For the focus interaction effect:

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

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

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

Combine all universities into one category

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

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

Name	Estimate	CI	pval
Tempt main effect within MTurk	0.21	[-0.28, 0.71]	0.40
Tempt main effect within university sites	0.28	[-0.08, 0.63]	0.12
Effect of university site vs. MTurk on tempt main effect	0.06	[-0.55, 0.67]	0.84
Tempt-load interaction within MTurk	-0.20	[-1.01, 0.60]	0.62
Tempt-load interaction within university sites	-0.17	[-0.69, 0.35]	0.52
Effect of university site vs. MTurk on tempt-load interaction	0.03	[-0.93, 0.99]	0.95

Refit original study's ANOVA model

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

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

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

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

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

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

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

Comparison to Many Labs 2

Questionnaire differences

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

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

Analysis differences

Additionally, the two studies used slightly different statistical analyses to aggregate data across sites: ML2 used an independent-samples t-test allowing for heteroskedasticity, while ML5 used a linear mixed model with model-based standard errors that assumed homoskedasticity. The original study used a one-way ANOVA (which assumes homoskedasticity). We therefore reanalyzed all subject-level data for primary analysis sites in both studies using identical statistical analyses using three approaches[2]:

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

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

1. To reproduce the analysis approach in ML5, we used a linear mixed model (**LMM**) to regress perceived likelihood on fixed effects of tempting fate, the interaction of tempting fate with study (ML5 vs. ML2), exchangeable random intercepts and slopes by site, and with normal errors. Inference used model-based standard errors, so assumed homoskedasticity.
2. To avoid parametric assumptions[3], we used generalized estimating equations (**GEE**) with a working independent correlation structure to regress perceived likelihood on tempting fate. Inference used robust standard errors, so made no assumptions about how subjects were correlated within sites or otherwise.
3. To reproduce the analysis approach in ML2, we computed raw mean differences (**RMD**) between tempting-fate conditions for each study separately, ignoring correlation of subjects within sites. We estimated standard errors for each as in the independent-samples Welch t-test conducted for ML2. We then tested the null hypothesis that the RMDs reflected the same population difference in both studies using the fact that, under the null:

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

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

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

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

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

Sampling frame differences

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

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

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

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

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

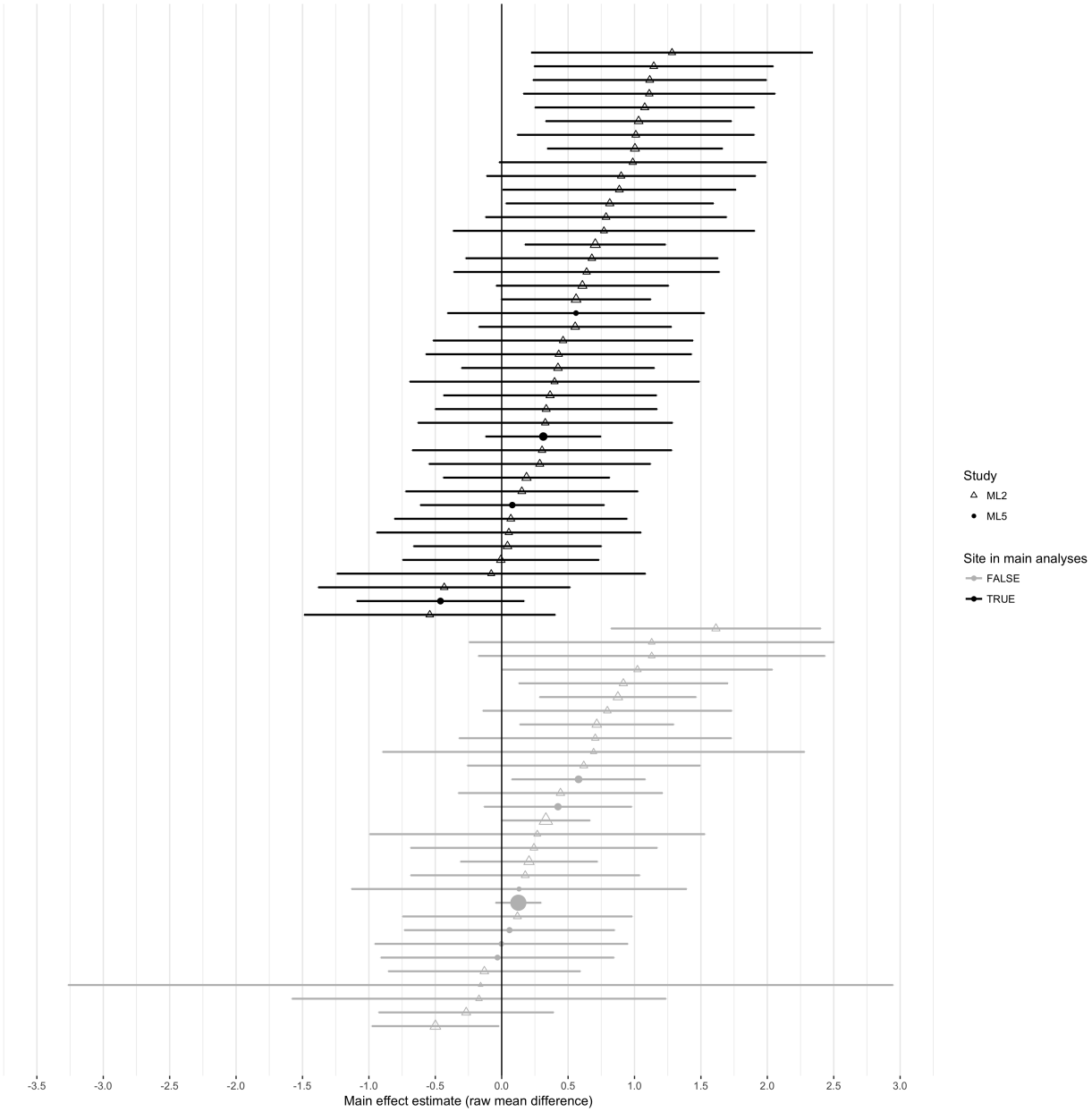
Combined estimates of tempting-fate effects

We combined all data from both studies to arrive at an updated estimate of the average effect of tempting fate across the diverse samples included in both studies. We again fit LMM and GEE models to primary analysis sites from both studies, omitting the interaction of tempting fate with study to estimate an average effect. The LMM model estimated a tempting-fate effect of 0.48 on the raw mean difference scale (95% CI: [0.34, 0.61]; $p = 1 \times 10^{-11}$). As expected, the GEE yielded a similar estimate (0.46) with more conservative inference (95% CI: [0.32, 0.61]; $p = 4 \times 10^{-10}$). The naïve standardized mean difference effect size was an estimated 0.19 (95% CI 0.14, 0.25). To assess consistency between this pooled estimate and the estimate of the original study, we estimated that, if the original study were consistent with all replications, then the probability of observing a point estimate in the original study at least as extreme as that actually observed would be approximately $P_{orig} = 0.28$ (Mathur and VanderWeele 2017). This method assumes normally distributed true effects, which appeared reasonable here.

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

Discussion

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



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

References

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

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