Online Supplement: Analysis Code

Contents

Data Quality

```
# total and excluded bad subjects
d = data.frame( site = first$site, n.excl = first$site.n.excl, n.total = first$site.n)

stargazer(d, header=FALSE, summary=FALSE,
          #column.labels = c("Site", "No. excluded subjects", "No. analyzed subjects"),
          rownames = FALSE,
          title="Excluded and analyzed subjects by site" )
```

Table 1: 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" )
```

Table 2: 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.

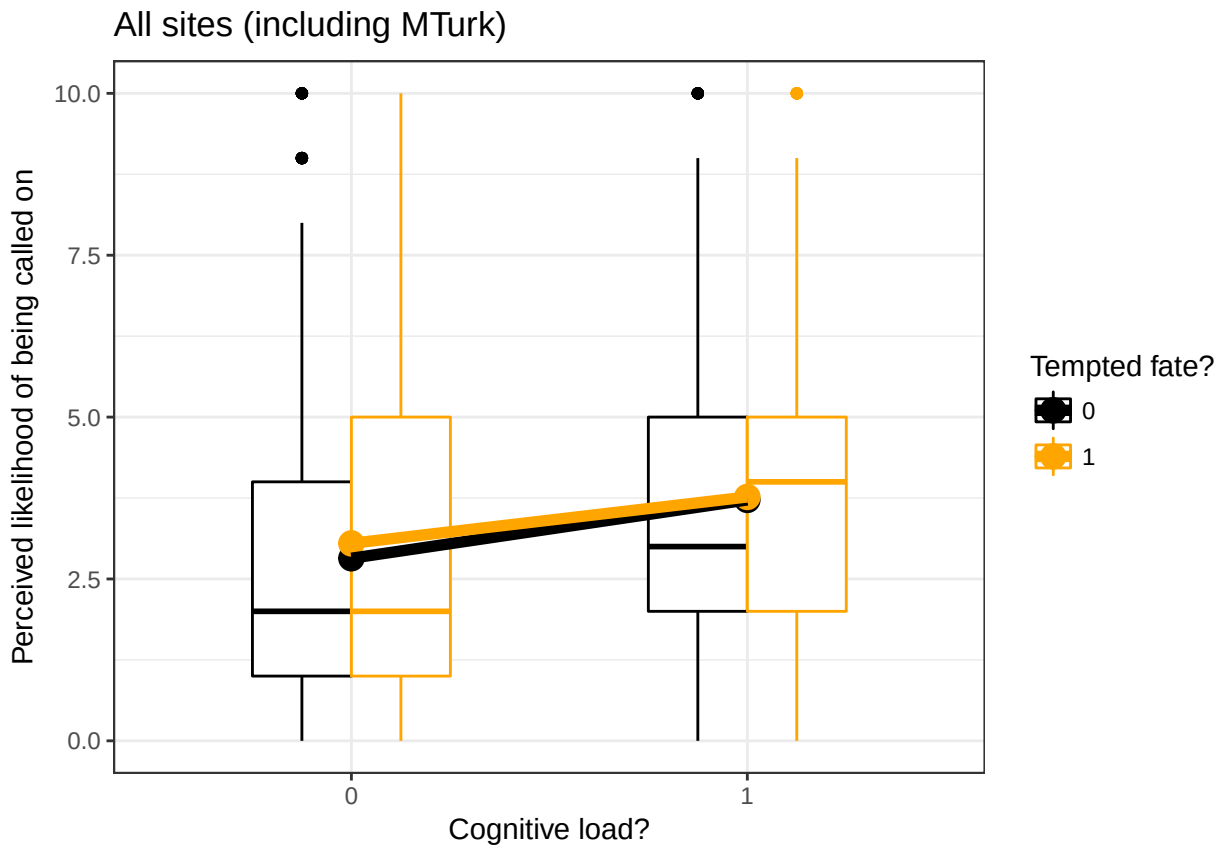
Descriptive Stats and Plots

Boxplots: medians and IQRs; lines: simple means by subset. (For the same plots within each site, see the data prep PDF.)

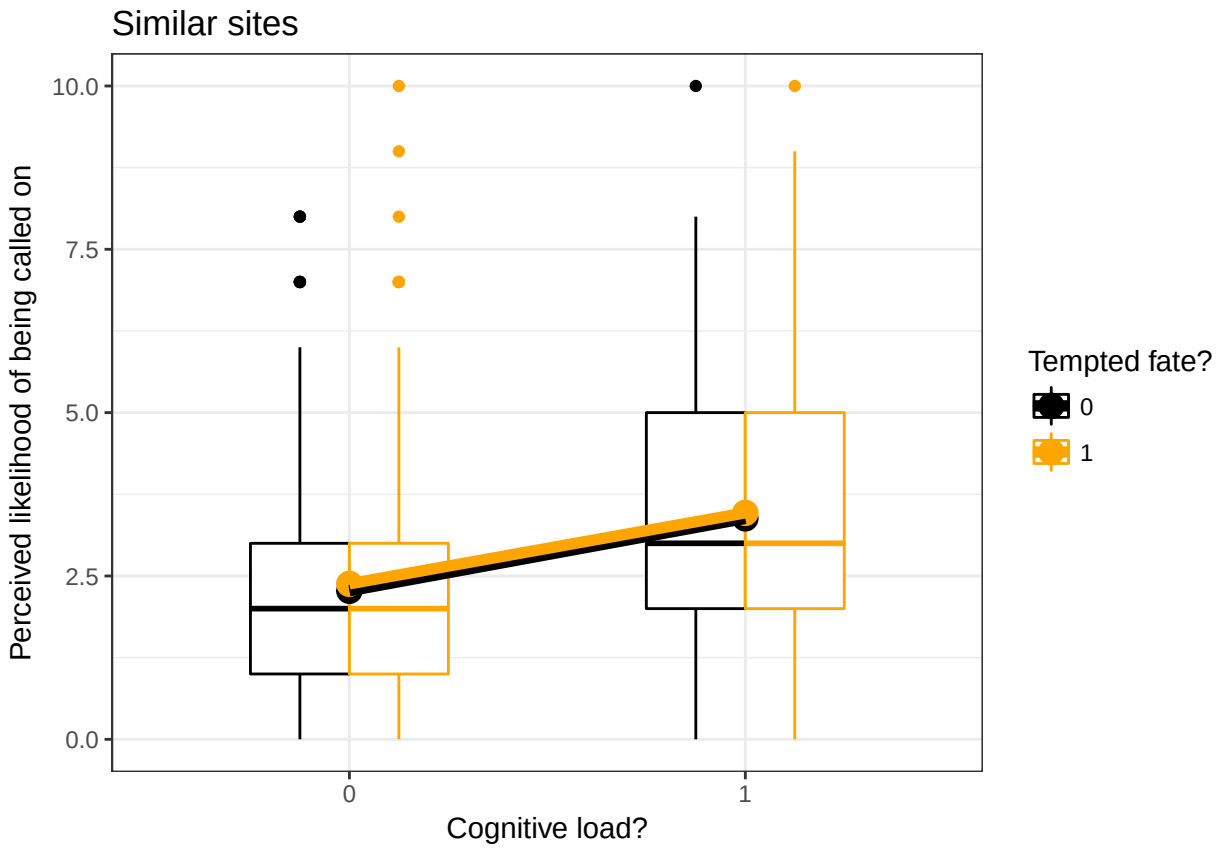
Note: These aggregated means and SDs pool across all sites within a group (similar, dissimilar, MTurk). We caution that such analyses are potentially subject to bias due to Simpson's Paradox (Rücker and Schumacher,

2008), which will be resolved in analysis models below by accounting for clustering by site. They are provided here only as descriptive summaries. The same caveat applies to the following section.

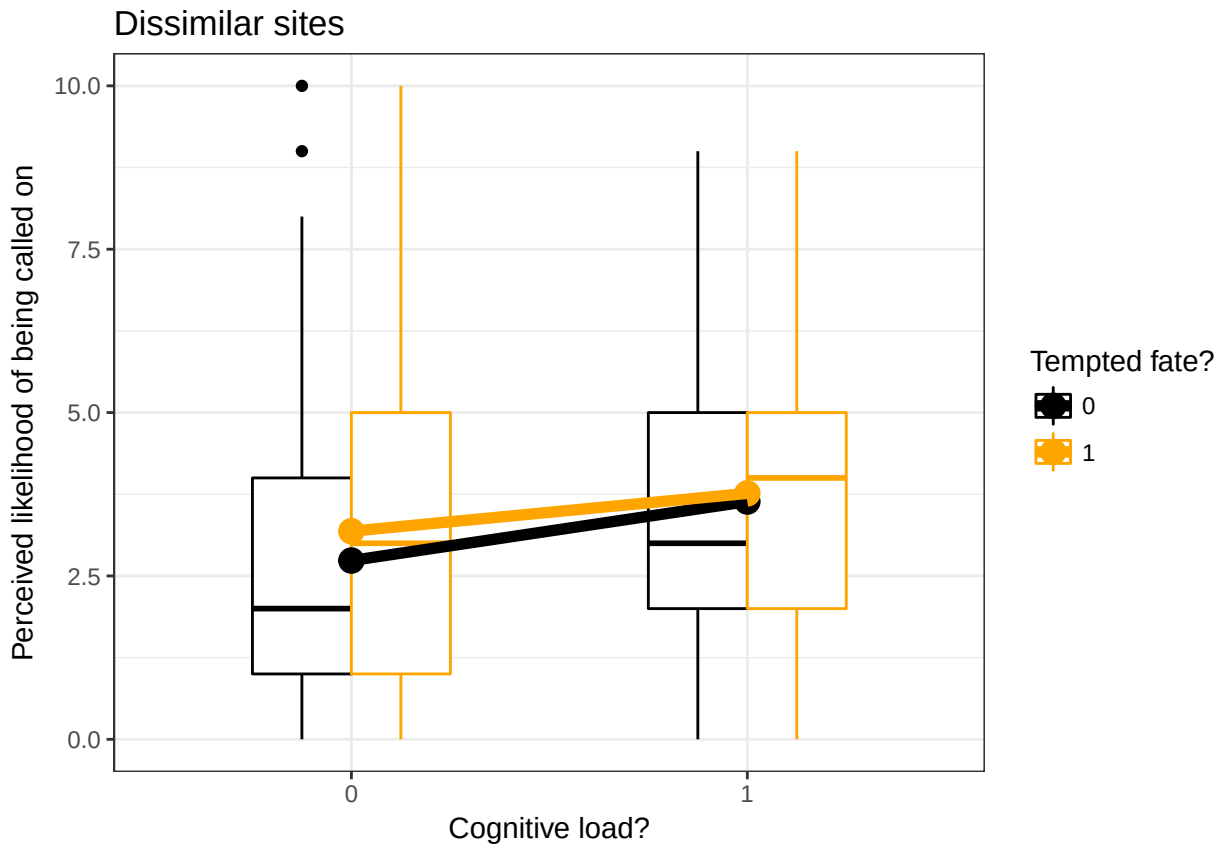
```
##### Fn: Interaction Plot #####
# pass the desired subset of data
int_plot = function( dat, ggtitle ) {
    agg = ddply( dat, .(load, tempt), summarize, val = mean(lkl, na.rm=TRUE)    )  # aggregate data for

  colors = c("black", "orange")
  plot( ggplot( dat, aes(x = as.factor(load), y = lkl, color=as.factor(tempt) ) ) + geom_boxplot(width=
      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
```



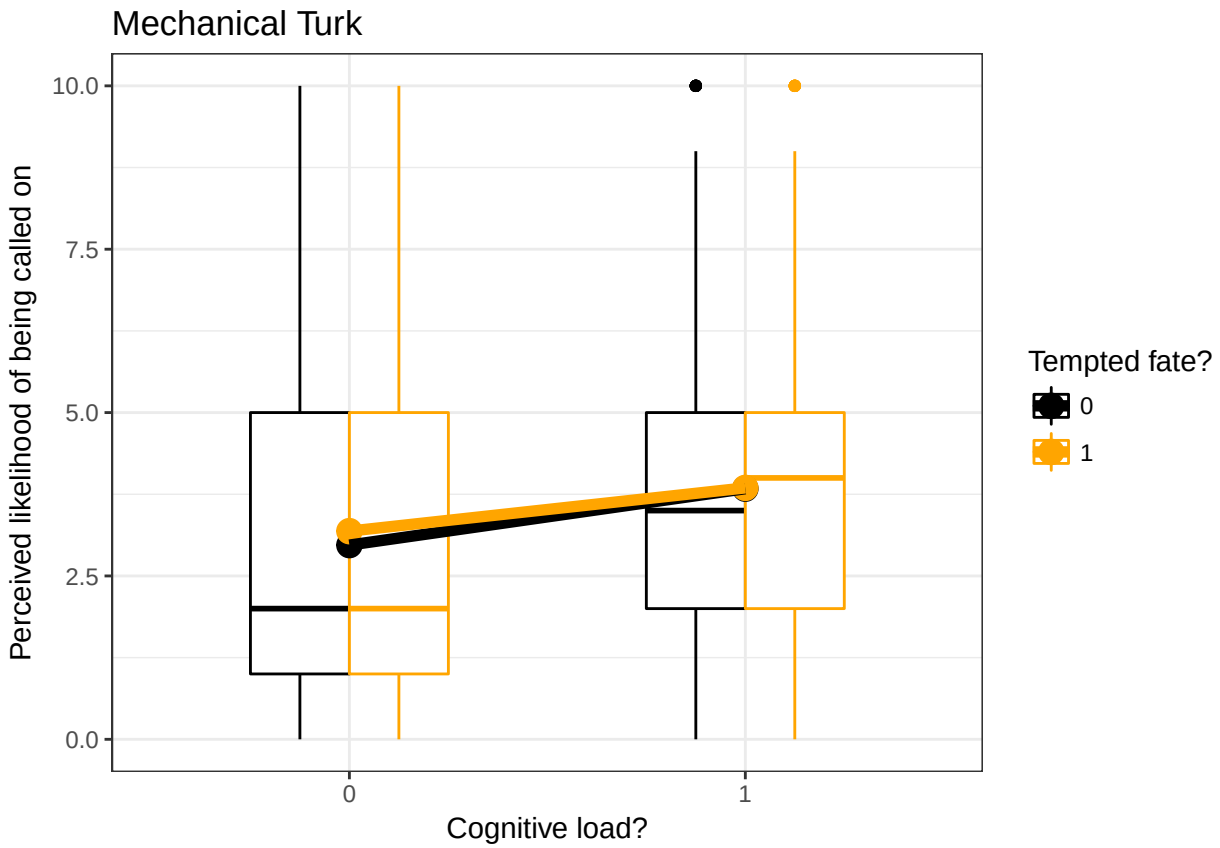

```
int_plot(b[ b$group=="b.similar", ], ggtitle = "Similar sites")
```




```
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)" )
```

Table 3: 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

Forest plots for main effect and interaction

Study-specific estimates are from OLS fit within just that site (this step was completed previously by `data_prep.Rmd`). Pooled estimates are based on estimated coefficients from LMMs (see preregistered protocol for exact model specification). Throughout, we use "main effect" to refer to the main effect in the condition without cognitive load.

(Technical note: An alternative for the study-specific estimates would be to use estimates of random intercepts and random slopes by site from the LMM, but here we use subset analyses for a descriptive characterization that relaxes the across-site distributional assumptions of LMM.)

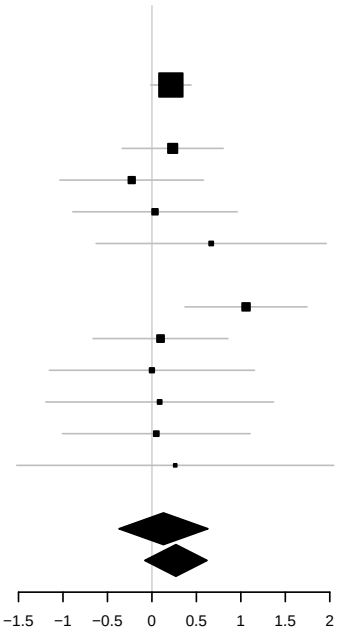
```
# first, fit models that we need for forest plot's pooled estimates
#  and subsequent analyses

# prevent brat forest plot from going off page
opts_chunk$set(echo=TRUE, tidy=TRUE, tidy.opts=list(width.cutoff=60), fig.width=10 )

# Fn: calculate SE for sum of coefficients
# b1, b2: names of the two coefficients to add
# .mod: the lmer model object
lin_combo = function( b1, b2, .mod ) {
  V = vcov(.mod)
  SE = sqrt( V[b1, b1] + V[b2, b2] + 2 * V[b1, b2] )
  est = fixef(.mod)[b1] + fixef(.mod)[b2]
  lo = as.numeric( est - qnorm(0.975) * SE )
  hi = as.numeric( est + qnorm(0.975) * SE )
  pval = ( 1 - pnorm( abs(est / SE) ) ) * 2

  return( data.frame( est, lo, hi, pval ) )
}



##### Only Similar Sites #####
# fit Primary Model 1, to be reported in subsequent section
```

```
# reference level for group is MTurk
m1 = lmer( lkl ~ tempt * load * group + (tempt * load | site), data = b[ b$group != "c.dissimilar", ] )

# bizarre mystery: changing order of variables in random slope specification
#  results in convergence failure:
# lmer( lkl ~ tempt * load * group + (load * tempt | site), data = b[ b$group != "c.dissimilar", ] )

# pooled estimate and CI of main effect (similar sites)
main.sim = lin_combo( "tempt", "tempt:groupb.similar", m1 )

# pooled estimate and CI of interaction (similar sites)
int.sim = lin_combo( "tempt:load", "tempt:load:groupb.similar", m1 )



##### Combining All Universities #####
# Model 1' in preregistered protocol
# here, reference level is all university sites
m2 = lmer( lkl ~ tempt * load * is.mturk + (tempt * load | site), data = b )

# pooled estimate and CI of main effect (all universities)
CI2 = confint(m2, method = "Wald")
main.uni = data.frame( est = fixef(m2)["tempt"], lo = CI2[ "tempt", 1 ],
                       hi = CI2[ "tempt", 2 ] )


# pooled estimate and CI of interaction (all universities)
int.uni = data.frame( est = fixef(m2)["tempt:load"], lo = CI2[ "tempt:load", 1 ],
                      hi = CI2[ "tempt:load", 2 ] )

# main effect in MTurk
mturk.main.m2 = lin_combo( "tempt:is.mturk", "tempt", m2 )

# interaction effect in MTurk
mturk.int.m2 = lin_combo( "tempt:load:is.mturk", "tempt:load", m2 )
```

```
# make the forest plot

# Fn: insert spacey elements in vectors for purely cosmetic
# forest plot reasons spaces are between site types 'use.NA'
# = should we put NA instead of ''?
pretty_spaces = function(x, use.NA = FALSE) {
    x2 = append(x, ifelse(use.NA, NA, ""), after = 1)
    x2 = append(x2, ifelse(use.NA, NA, ""), after = 6)
}


# build text 'columns' of forest plot NAs are for making
# spaces
tabletext.main = cbind(c("Study", "", pretty_spaces(as.character(first$site)),
    NA, "Pooled (similar universities)", "Pooled (all universities)"),
    c("Sample size", "", pretty_spaces(first$site.n), NA, sum(first$site.n[first$group ==
        "b.similar"]), sum(first$site.n[!first$is.mturk])), c("Main effect est.",
        "", pretty_spaces(round(first$site.main.est, 2)), NA,
```

```
        round(main.sim$est, 2), round(main.uni$est, 2)))

# build columns of point estimates, CI lower, and CI upper
# values for forest plot
m.main = c(NA, NA, pretty_spaces(first$site.main.est, use.NA = TRUE),
    NA, round(main.sim$est, 2), round(main.uni$est, 2))
l.main = c(NA, NA, pretty_spaces(first$site.main.lo, use.NA = TRUE),
    NA, round(main.sim$lo, 2), round(main.uni$lo, 2))
u.main = c(NA, NA, pretty_spaces(first$site.main.hi, use.NA = TRUE),
    NA, round(main.sim$hi, 2), round(main.uni$hi, 2))


forestplot(labeltext = tabletext.main, mean = m.main, lower = l.main,
    upper = u.main, zero = 0, is.summary = c(TRUE, rep(FALSE,
        14), TRUE, TRUE))
# lines( x = c( 1.54, 1.54 ), y = c( 0.09205192, 0.863934 ),
# lty = 2, col = 'red' )
yi.orig.main = 2.93 - 1.9
abline(v = yi.orig.main, lty = 2, col = "red")
```

Study	Sample size	Main effect est.
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)	**1477**	**0.27**

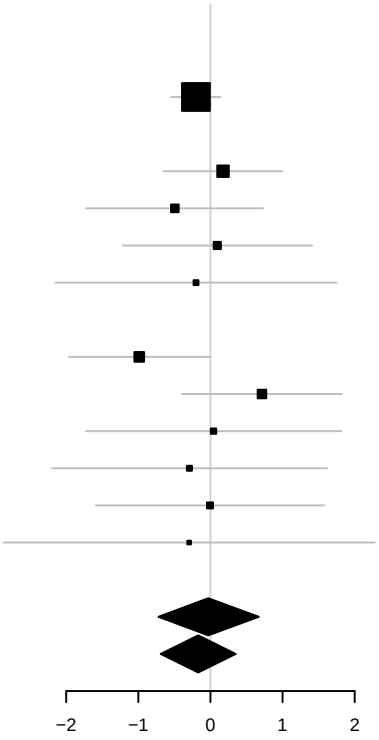



```
# text( yi.orig.main, .4, 'Original study', col = 'red' )
```

```
######## For interaction ########

# build text 'columns' of forest plot NAs are for making
# spaces
tabletext.int = cbind(c("Study", "", pretty_spaces(as.character(first$site)),
    NA, "Pooled (similar universities)", "Pooled (all universities)"),
    c("Sample size", "", pretty_spaces(first$site.n), NA, sum(first$site.n[first$group ==
        "b.similar"]), sum(first$site.n[!first$is.mturk])), c("Interaction est.",
        "", pretty_spaces(round(first$site.int.est, 2)), NA,
        round(int.sim$est, 2), round(int.uni$est, 2)))
```

```r
# build columns of point estimates, CI lower, and CI upper
# values for forest plot
m.int = c(NA, NA, pretty_spaces(first$site.int.est, use.NA = TRUE),
    NA, round(int.sim$est, 2), round(int.uni$est, 2))
l.int = c(NA, NA, pretty_spaces(first$site.int.lo, use.NA = TRUE),
    NA, round(int.sim$lo, 2), round(int.uni$lo, 2))
u.int = c(NA, NA, pretty_spaces(first$site.int.hi, use.NA = TRUE),
    NA, round(int.sim$hi, 2), round(int.uni$hi, 2))

forestplot(labeltext = tabletext.int, mean = m.int, lower = l.int,
    upper = u.int, zero = 0, is.summary = c(TRUE, rep(FALSE,
        14), TRUE, TRUE))
```

Study	Sample size	Interaction est.
MTurk	2973	−0.2
U Penn	335	0.17
UCB	200	−0.49
UVA	160	0.09
Stanford	68	−0.2
Eotvos	284	−0.99
KUL	118	0.71
PUC	91	0.04
BYU	84	−0.29
URI	81	−0.01
RHIT	56	−0.3
Pooled (similar universities)	**763**	**−0.03**
Pooled (all universities)	**1477**	**−0.17**



Sanity check: Estimates in the forest plots seem to agree closely with the interaction plots by site type as well as interaction plots for each site (`data_prep.pdf`).

Planned Primary Analyses

Model 1: Observation-level mixed model

Model 1 is a linear mixed model excluding dissimilar sites. We use X to denote tempting fate, L to denote cognitive load, and Y to denote perceived likelihood.

```r
# see section before making forest plot for model fit note
# that reference level for site type is MTurk

# using Wald CIs because profile and boot are struggling to
```

```
# converge (i.e., assume coefficient estimates are normal,
# which is quite reasonable at these sample sizes)
CI = confint(m1, method = "Wald")

# make table
name = c("Magnitude of X main effect within MTurk", "Magnitude of X main effect within similar sites",
    "Effect of similar site vs. MTurk on X main effect", "Magnitude of X-L interaction within MTurk",
    "Magnitude of X-L interaction within similar sites", "Effect of similar site vs. MTurk on X-L intera

value = as.numeric(c(fixef(m1)["tempt"], main.sim$est, fixef(m1)["tempt:groupb.similar"],
    fixef(m1)["tempt:load"], int.sim$est, fixef(m1)["tempt:load:groupb.similar"]))
value = round(value, 2)

lo = as.numeric(c(CI["tempt", 1], main.sim$lo, CI["tempt:groupb.similar",
    1], CI["tempt:load", 1], int.sim$lo, CI["tempt:load:groupb.similar",
    1]))
lo = round(lo, 2)

hi = as.numeric(c(CI["tempt", 2], main.sim$hi, CI["tempt:groupb.similar",
    2], CI["tempt:load", 2], int.sim$hi, CI["tempt:load:groupb.similar",
    2]))
hi = round(hi, 2)

CI.string = paste("[", lo, ", ", hi, "]", sep = "")

pvals.m1 = coef(summary(m1))[, 5]
pval = as.numeric(c(pvals.m1["tempt"], main.sim$pval, pvals.m1["tempt:groupb.similar"],
    pvals.m1["tempt:load"], int.sim$pval, pvals.m1["tempt:load:groupb.similar"]))
pval = round(pval, 2)

m1.res = data.frame(Name = name, Estimate = value, CI = CI.string,
    pval = pval)
kable(m1.res)
```

Name	Estimate	CI	pval
Magnitude of X main effect within MTurk	0.21	[-0.27, 0.7]	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

Sanity check: Instead of fitting 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.

Planned Secondary Analyses

Model 1': Observation-level mixed model, including dissimilar sites

We refit the primary analysis model, but now including the dissimilar sites. (This model was actually already fit for the pooled estimate in the forest plots.)

```
# make table
name = c("Magnitude of X main effect within MTurk", "Magnitude of X main effect within university sites
    "Effect of university site vs. MTurk on X main effect", "Magnitude of X-L interaction within MTurk"
    "Magnitude of X-L interaction within university sites", "Effect of university site vs. MTurk on X-L

# negative ones are when coefficient is ( MTurk - uni )
value = as.numeric(c(mturk.main.m2$est, fixef(m2)["tempt"], -fixef(m2)["tempt:is.mturk"],
    mturk.int.m2$est, fixef(m2)["tempt:load"], -fixef(m2)["tempt:load:is.mturk"]))
value = round(value, 2)

lo = as.numeric(c(mturk.main.m2$lo, CI2[row.names(CI2) == "tempt",
    1], -CI2[row.names(CI2) == "tempt:is.mturk", 1], mturk.int.m2$lo,
    CI2[row.names(CI2) == "tempt:load", 1], -CI2[row.names(CI2) ==
        "tempt:load:is.mturk", 1]))
lo = round(lo, 2)

hi = as.numeric(c(mturk.main.m2$hi, CI2[row.names(CI2) == "tempt",
    2], -CI2[row.names(CI2) == "tempt:is.mturk", 2], mturk.int.m2$hi,
    CI2[row.names(CI2) == "tempt:load", 2], -CI2[row.names(CI2) ==
        "tempt:load:is.mturk", 2]))
hi = round(hi, 2)

CI.string = paste("[", lo, ", ", hi, "]", sep = "")

pvals.m2 = coef(summary(m2))[, 5]
pval = as.numeric(c(mturk.main.m2$pval, pvals.m2["tempt"], pvals.m2["tempt:is.mturk"],
    mturk.int.m2$pval, pvals.m2["tempt:load"], pvals.m2["tempt:load:is.mturk"]))
pval = round(pval, 2)

m2.res = data.frame(Name = name, Estimate = value, CI = CI.string,
    pval = pval)
kable(m2.res)
```

Name	Estimate	CI	pval
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, 0.6]	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

As a sanity check, work in the "statistical consistency" sections below demonstrates that meta-analytic counterparts to these observation-level models yield nearly identical results.

Model 2: Moderation by median SAT score

We treated university sites' median total SAT scores (estimated for 2018) as a proxy for similarity to the site of the original study (Cornell), assuming that universities with higher SAT scores are more similar to Cornell (median SAT: 2134). Universities outside the US and MTurk were given missing values for SAT score. Model 2 assesses whether median SAT score moderates the effect of interest.

```
# center and scale SAT
b$SATc = (b$SAT - mean(b$SAT, na.rm = TRUE))/sd(b$SAT, na.rm = TRUE)

m.sat = lmer(lkl ~ tempt * load * SATc + (tempt * load | site),
    data = b)

CI.SAT = confint(m.sat, method = "Wald")

summary(m.sat)
```

```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## Formula: lkl ~ tempt * load * SATc + (tempt * load | site)
##    Data: b
##
## REML criterion at convergence: 4212.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8762 -0.7428 -0.2423  0.7119  3.6645
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  site     (Intercept) 0.04342  0.2084
##           tempt       0.02762  0.1662   -0.99
##           load        0.05230  0.2287    0.70 -0.80
##           tempt:load  0.04249  0.2061   -1.00  0.99 -0.69
##  Residual             4.17002  2.0421
## Number of obs: 984, groups:  site, 7
##
## Fixed effects:
##                 Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)      2.42197    0.15407  5.91000  15.720 4.78e-06 ***
## tempt            0.10888    0.19078 14.54000   0.571  0.57694
## load             1.01817    0.21028  8.38000   4.842  0.00113 **
## SATc            -0.20542    0.14220 12.10000  -1.445  0.17397
## tempt:load      -0.07719    0.27555 19.34000  -0.280  0.78235
## tempt:SATc       0.03811    0.18334 34.84000   0.208  0.83652
## load:SATc        0.13426    0.20295 20.86000   0.662  0.51549
## tempt:load:SATc  0.01044    0.27065 50.17000   0.039  0.96938
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Correlation of Fixed Effects:
##             (Intr) tempt  load   SATc   tmpt:l tm:SAT ld:SAT
## tempt       -0.751
## load        -0.319  0.269
## SATc         0.091 -0.058  0.040
## tempt:load   0.197 -0.486 -0.697 -0.044
## tempt:SATc  -0.056  0.041 -0.028 -0.739  0.023
## load:SATc    0.039 -0.029  0.046 -0.426 -0.017  0.338
## tmpt:ld:SAT -0.042  0.023 -0.016  0.291  0.024 -0.549 -0.710
```

This analysis does not support moderation of either the main effect or the interaction by median SAT score.

Refitting original 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. However, we cautio that unlike our primary model, the present analysis that does not account for site is potentially subject to bias due to Simpson's Paradox.

```
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.

Post-Hoc Analyses

Statistical consistency of main effect estimates between original and replications (similar sites only)

The original study's Experiment 6 reported (for the no-load condition only):

- $\bar{Y}_{X=0,L=0} = 1.90, SD_{Y=0,L=0} = 1.42, n = 30$
- $\bar{Y}_{X=1,L=0} = 2.93, SD_{Y=1,L=0} = 2.16, n = 30$

```
# effect sizes of original
yi.orig.main = 2.93 - 1.9
var.mean0 = 1.42^2/30
var.mean1 = 2.16^2/30
vyi.orig.main = var.mean0 + var.mean1

# sanity check: try to reproduce t-stat in original paper
(SMD.orig.main = yi.orig.main/sqrt(vyi.orig.main))
```

```
## [1] 2.182452
```

```
# matches their t= 2.19 (pg 302, column 2)

#
```

We next estimate the mean and heterogeneity of the site-specific effects among the replications using a mixed model similar to Model 1, except using only similar sites (not MTurk). Note that since these analyses only use the 4 similar sites, heterogeneity estimation is likely to be pretty unstable.

```
detach("package:lmerTest")
# detach('package:nlme')
m = lmer(lkl ~ tempt * load + (tempt * load | site), data = b[b$group ==
    "b.similar", ])
Vhat = 0.05701  # variance of random slopes of tempt
Mhat = fixef(m)[["tempt"]]
SE.Mhat = sqrt(vcov(m)["tempt", "tempt"])
```

Compute P_{orig}, i.e., the probability that the original estimate would be as extreme or more extreme than it actually was if drawn from the estimated effect distribution from the replications (Mathur and VanderWeele, 2017):

```
(p.orig.main.sim = p_orig(orig.y = yi.orig.main, orig.vy = vyi.orig.main,
    yr = Mhat, t2 = Vhat, vyr = SE.Mhat^2))
```

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

## [1] 0.1206834
```

Sanity check: Try meta-analyzing the sites' point estimates instead.

```
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
```

It's a bit lower due to the lower estimated heterogeneity here.

Sanity check: Do these results agree with prediction intervals? They should because there is still basically zero heterogeneity.

```
pred = pred_int(orig.y = yi.orig.main, orig.vy = vyi.orig.main,
    rep.y = first[first$group == "b.similar", ]$site.main.est,
    rep.vy = first[first$group == "b.similar", ]$site.main.SE^2)
```

3 of 4 similar sites are within their prediction intervals.

Statistical consistency of interaction estimates between original and replications (similar sites only)

```
# interaction is the 'difference in differences'
yi.orig.int = (5.27 - 2.7) - (2.93 - 1.9)
vyi.orig.int = (1.42^2/30) + (2.16^2/30) + (2.17^2/30) + (2.36^2/30)
# just add the variances that contribute to the linear combo

# sanity check: reproduce original paper's F-stat
SMD.orig.int = yi.orig.int/sqrt(vyi.orig.int)
SMD.orig.int^2  # square a t-stat
```

```
## [1] 4.194923
```

```
# appears within rounding error (reported: F = 4.15)
```

We again estimate the mean and heterogeneity of the site-specific effects among the replications using the same mixed model (among only similar sites) that we fit above.

```
# same mixed model as above
Vhat = 0.14362  # variance of random slopes of tempt:load
Mhat = fixef(m)[["tempt:load"]]
SE.Mhat = sqrt(vcov(m)["tempt:load", "tempt:load"])
```

Compute P_{orig}:

```
p.orig.int.sim = p_orig(orig.y = yi.orig.int, orig.vy = vyi.orig.int,
    yr = Mhat, t2 = Vhat, vyr = SE.Mhat^2)
```

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

Sanity check: Use meta-analysis instead.

```
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
```

```
pred = pred_int(orig.y = yi.orig.int, orig.vy = vyi.orig.int,
    rep.y = first[first$group == "b.similar", ]$site.int.est,
    rep.vy = first[first$group == "b.similar", ]$site.int.SE^2)
```

3 of 4 similar sites are within their prediction intervals. This seems reasonable given P_{orig}.

Statistical consistency of main effect estimates between original and replications (all university sites)

We now consider consistency of the original study with all university replications. This allows for more precise estimation of heterogeneity.

Fit a mixed model excluding only MTurk:

```
# detach('package:lmerTest') detach('package:nlme')
m = lmer(lkl ~ tempt * load + (tempt * load | site), data = b[!b$is.mturk,
    ])
Vhat = 0.06692  # variance of random slopes of tempt; manual because extracting the object is huge pain
Mhat = fixef(m)[["tempt"]]
SE.Mhat = sqrt(vcov(m)["tempt", "tempt"])
```

Compute P_{orig}:

```
(p.orig.main.uni = p_orig(orig.y = yi.orig.main, orig.vy = vyi.orig.main,
    yr = Mhat, t2 = Vhat, vyr = SE.Mhat^2))
```

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

## [1] 0.1766483
```

As a sanity check, try meta-analyzing the sites' point estimates instead:

```
meta.int = rma.uni(yi = site.main.est, vi = site.main.SE^2, data = first[!first$is.mturk,
    ], 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
```

The estimated main effect and heterogeneity in similar sites was $\widehat{M} = 0.27$ and $\widehat{V} = 0.07$ in the mixed model compared to $\widehat{M} = 0.13$ and $\widehat{V} = 0$ in the meta-analysis. They agree very closely.

Another sanity check: Do these results, suggestive of good consistency, agree with prediction intervals? They should because there is basically zero heterogeneity.

```
pred = pred_int(orig.y = yi.orig.main, orig.vy = vyi.orig.main,
    rep.y = first[!first$is.mturk, ]$site.main.est, rep.vy = first[!first$is.mturk,
        ]$site.main.SE^2)
```

9 of 10 university sites are within their prediction intervals. This is close to the 95% we would expect under consistency.

Statistical consistency of interaction estimates between original and replications (all university sites)

We again estimate the mean and heterogeneity of the site-specific effects among the replications using the same mixed model (among only similar sites) that we fit above.

```
# same mixed model as above
Vhat = 0.05823  # variance of random slopes of tempt:load
Mhat = fixef(m)[["tempt:load"]]
SE.Mhat = sqrt(vcov(m)["tempt:load", "tempt:load"])
```

Compute P_{orig}:

```
(p.orig.int.uni = p_orig(orig.y = yi.orig.int, orig.vy = vyi.orig.int,
    yr = Mhat, t2 = Vhat, vyr = SE.Mhat^2))
```

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

## [1] 0.03838219
```

As a sensitivity analysis, use meta-analysis instead:

```
meta.int = rma.uni(yi = site.int.est, vi = site.int.SE^2, data = first[!first$is.mturk,
    ], 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] 0.03491295
```

The estimated interaction and heterogeneity in similar sites was $\widehat{M}$ = -0.17 and $\widehat{V}$ = 0.06 in the mixed model compared to $\widehat{M}$ = -0.11 and $\widehat{V}$ = 0 in the meta-analysis. They agree very closely.

Sanity check: Do these relatively poor consistency results agree with prediction intervals?

Sanity check: Do these poor consistency results agree with prediction intervals?

```
pred = pred_int(orig.y = yi.orig.int, orig.vy = vyi.orig.int,
    rep.y = first[!first$is.mturk, ]$site.int.est, rep.vy = first[!first$is.mturk,
        ]$site.int.SE^2)
```

8 of 10 university sites are within their prediction intervals. This is borderline compared to expectation, as is P_{orig} when compared to the corresponding $\alpha = 0.05$ threshold.

Effectiveness of cognitive load manipulation on MTurk

Is the cognitive load manipulation less effective in MTurk vs. all universities combined? That is, does its effect on the tempt * load interaction vary between MTurk and all universities combined?

```
(m.manip = lmer(lkl ~ tempt * load * is.mturk + (tempt * load |
    site), data = b))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: lkl ~ tempt * load * is.mturk + (tempt * load | site)
##    Data: b
## REML criterion at convergence: 19902.43
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  site     (Intercept) 0.4260
##           tempt       0.2271    0.47
##           load        0.2887    0.86  0.41
##           tempt:load  0.3753   -0.96 -0.68 -0.88
##  Residual             2.2554
## Number of obs: 4450, groups:  site, 11
## Fixed Effects:
##         (Intercept)                tempt                 load
##             2.50293              0.27354              1.00020
##            is.mturk           tempt:load       tempt:is.mturk
##             0.46920             -0.17005             -0.06025
##       load:is.mturk  tempt:load:is.mturk
##            -0.13694             -0.03119
```

```
# mturk cannot have its own random slope because only one
# such site
CI.manip = confint(m.manip, method = "Wald")
```

- Effect of MTurk vs. university on effect of cognitive load ($L * Turk$) interaction: -0.03, 95% CI: -0.99, 0.93.

Are subjects' reported difficulty or effort associated with the cognitive load manipulation less for MTurk vs. all universities combined?

```
# subset to only subjects actually assigned to cognitive load
# mturk cannot have its own random slope because only one
# such site
(m.effort = lmer(count.eff ~ is.mturk + (1 | site), data = b[b$load ==
    1, ]))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: count.eff ~ is.mturk + (1 | site)
##    Data: b[b$load == 1, ]
## REML criterion at convergence: 7840.761
## Random effects:
##  Groups   Name        Std.Dev.
##  site     (Intercept) 0.4969
##  Residual             1.9890
## Number of obs: 1857, groups:  site, 10
## Fixed Effects:
## (Intercept)     is.mturk
```

```
##      7.1445       0.6235
```

```
CI.effort = confint(m.effort, method = "Wald")

(m.hard = lmer(count.hard ~ is.mturk + (1 | site), data = b[b$load ==
    1, ]))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: count.hard ~ is.mturk + (1 | site)
##    Data: b[b$load == 1, ]
## REML criterion at convergence: 8247.211
## Random effects:
##  Groups   Name        Std.Dev.
##  site     (Intercept) 0.2804
##  Residual             2.2340
## Number of obs: 1853, groups:  site, 10
## Fixed Effects:
## (Intercept)     is.mturk
##      6.7315       0.5038
```

```
# cannot include random slopes as random slopes due to
# convergence problems
CI.hard = confint(m.hard, method = "Wald")
```

- Effect of MTurk vs. university on perceived effort needed for cognitive load task: 0.62, 95% CI: -0.43, 1.67
- Effect of MTurk vs. university on perceived difficulty of cognitive load task: 0.5, 95% CI: -0.12, 1.13

Summary: There is no evidence here that the cognitive load manipulation is less effective on MTurk than in universities, either based on its actual effect on likelihood judgements or on its subjective impact.

More on MTurk vs. college students

How much do students care about answering questions correctly in class by site?

```
kable(aggregate(importance ~ group, FUN = mean, data = b))
```

group	importance
a.mturk	7.402334
b.similar	6.395018
c.dissimilar	6.716502

```
kable(aggregate(badness ~ group, FUN = mean, data = b))
```

group	badness
a.mturk	7.368385
b.similar	7.412811
c.dissimilar	6.968970

```r
summary(lm((b$importance - mean(b$importance, na.rm = TRUE)) ~
    site, data = b))
```

```
##
## Call:
## lm(formula = (b$importance - mean(b$importance, na.rm = TRUE)) ~
##     site, data = b)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4023 -1.4023  0.5977  1.5977  4.1964
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.24605    0.24855  -0.990   0.3223
## siteEotvos    0.05637    0.28304   0.199   0.8422
## siteKUL      -0.32002    0.32519  -0.984   0.3251
## siteMTurk     0.49757    0.25210   1.974   0.0485 *
## sitePUC      -0.48903    0.34652  -1.411   0.1582
## siteRHIT     -1.10119    0.39299  -2.802   0.0051 **
## siteStanford -0.69581    0.37313  -1.865   0.0623 .
## siteU Penn   -0.51969    0.27797  -1.870   0.0616 .
## siteURI      -0.08198    0.35702  -0.230   0.8184
## siteUVA      -0.41101    0.30693  -1.339   0.1806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.278 on 4174 degrees of freedom
##   (266 observations deleted due to missingness)
## Multiple R-squared:  0.0322, Adjusted R-squared:  0.03011
## F-statistic: 15.43 on 9 and 4174 DF,  p-value: < 2.2e-16
```

```r
summary(lm((b$badness - mean(b$importance, na.rm = TRUE)) ~ site,
    data = b))
```

```
##
## Call:
## lm(formula = (b$badness - mean(b$importance, na.rm = TRUE)) ~
##     site, data = b)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5075 -1.3684  0.6316  1.6316  4.0000
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.4127     0.2699  -1.529   0.1263
## siteEotvos     0.4209     0.3074   1.369   0.1710
## siteKUL        0.3636     0.3532   1.030   0.3033
## siteMTurk      0.6303     0.2738   2.302   0.0214 *
## sitePUC        0.1383     0.3763   0.368   0.7132
## siteRHIT      -0.7381     0.4268  -1.729   0.0838 .
## siteStanford   0.4559     0.4052   1.125   0.2606
## siteU Penn     0.7694     0.3019   2.549   0.0108 *
```

```
## siteURI        0.3885     0.3877   1.002   0.3164
## siteUVA        0.5682     0.3333   1.704   0.0884 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.474 on 4171 degrees of freedom
##   (269 observations deleted due to missingness)
## Multiple R-squared:  0.006939,   Adjusted R-squared:  0.004797
## F-statistic: 3.238 on 9 and 4171 DF,  p-value: 0.0006308
```

Do MTurkers care less than students?

```
summary(lm((b$importance - mean(b$importance, na.rm = TRUE)) ~
    is.mturk, data = b))
```

```
##
## Call:
## lm(formula = (b$importance - mean(b$importance, na.rm = TRUE)) ~
##     is.mturk, data = b)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4023 -1.4023  0.5977  1.5977  3.4256
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5765     0.0640  -9.007   <2e-16 ***
## is.mturk      0.8280     0.0767  10.794   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.282 on 4182 degrees of freedom
##   (266 observations deleted due to missingness)
## Multiple R-squared:  0.02711,    Adjusted R-squared:  0.02687
## F-statistic: 116.5 on 1 and 4182 DF,  p-value: < 2.2e-16
```

```
summary(lm((b$badness - mean(b$importance, na.rm = TRUE)) ~ is.mturk,
    data = b))
```

```
##
## Call:
## lm(formula = (b$badness - mean(b$importance, na.rm = TRUE)) ~
##     is.mturk, data = b)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3684 -1.3684  0.6316  1.8348  2.8348
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01441    0.06952   0.207   0.8358
## is.mturk     0.20316    0.08333   2.438   0.0148 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.478 on 4179 degrees of freedom
##   (269 observations deleted due to missingness)
## Multiple R-squared:  0.00142,    Adjusted R-squared:  0.001181
## F-statistic: 5.944 on 1 and 4179 DF,  p-value: 0.01481
```

Actually, they care more.

Challenges in replicating Risen & Gilovich (2008): a registered multisite replication

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

- Table 1 is in the file "Table 1. csv". Format and submit it as separate document because it's very wide.
- Say "original" and "updated" protocol (not "endorsed" / "non-endorsed")
- Add vertical line in forest plots for original estimates
- One site still hasn't filled in its protocol characteristics
- Fix number of digits throughout
- Remove Simpson's Paradox because binary variables?
- Check w/ UCB about prior task

Rules:

"Given that we'll have 11 of these, I'd like to keep them as concise as possible. I definitely want to avoid having each paper discuss the motivation for the overall project (e.g., the idea of compare a reviewed protocol to the original one, the lack of full vetting/acceptance of the RP:P protocols, the drawbacks of the one-off RP:P approach relative to this more RRR-like approach, etc.). Those general issues should appear only in the main overview paper so that we don't have 11 papers making the same points. My hope would be to have the finding-specific papers keep the introduction really short, perhaps only a couple of paragraphs. The method and results should be complete enough to convey the study procedures and results fully. Perhaps it would work to refer readers to the OSF project page and the original RP:P materials for details of the original RP:P protocol, with the ML5 papers thoroughly describing the differences between the protocols and the characteristics of the new sample and testing. So, keep the intro and discussions sections for these individual finding paper short (no need for extensive theorizing or overviews) and make the method/results complete enough."

Abstract

WC: 248/250

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 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 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 (CITE). Thus, Risen et al. (2008) additionally hypothesized that System 2 processes can act to suppress irrational aversion to 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 perceived likelihood of a 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 both a main effect (standardized mean difference [SMD] = 2.18, F = XX, p = XX)[1] and the target interaction effect (SMD = 2.18, F = XX, p = XX).

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 (standardized mean difference [SMD] = 0.5, F(1,222) = 0.50, p = 0.48) or the target interaction (SMD = 0.05, F(1,222) = 0.002, 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 authors felt that the experimental scenario regarding being unprepared to answer questions in class may be less personally salient to subjects not enrolled in an elite university similar to Cornell, where the original study was conducted. Talk about ML2 results. Thus, the present multisite replication project aimed to: (1) reassess replicability of (Risen et al., 2008) using an updated protocol designed in collaboration with Risen and Gilovich to mitigate potential problems with the previous replication protocol; and (2) formally assess the effect of the updated protocol by comparing its results to newly collected results under the previous replication protocol.

Methods

The protocol, sample size criteria, and statistical analysis plan were preregistered[2] with details publicly available (CITE); any departures from these plans are reported in this manuscript. We designed the updated protocol in collaboration with the original authors (JR) and editor Daniel Simons (DS), 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 only in physical 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

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

[2] One site (BYU) was permitted to collect data prior to preregistration of the 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.

larger room, but in individual cubicles to minimize social distractions. To minimize potential contamination, sites wishing to run unrelated experiments on the same subjects prior to their participation in the present experiment were required to submit these tasks for approval by MM, JR, and DS; in practice, say how many sites actually did this. For sites whose subjects were not expected to speak fluent English, questionnaire materials were translated and verified through independent back-translation.

We additionally used the previous RPP protocol without modification to collect a new sample on Amazon Mechanical Turk ("MTurk"). Finally, we collected secondary data in several universities that did not meet the geographic or 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.

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.

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) to primary analyses; the MTurk sample contributed $n = 2973$ analyzed subjects to primary analyses. An additional six 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 a separate 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. 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 was almost identical to the estimate previously obtained under the same protocol in RPP (0.21 in the present sample vs. 0.2 in RPP). Considering all 10 university sites, 9 had main effect estimates in the same

direction as the original study. However, all of 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 study estimate, 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 RPP 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; $p = 0.05$).

Replication results under the updated protocol

bookmark

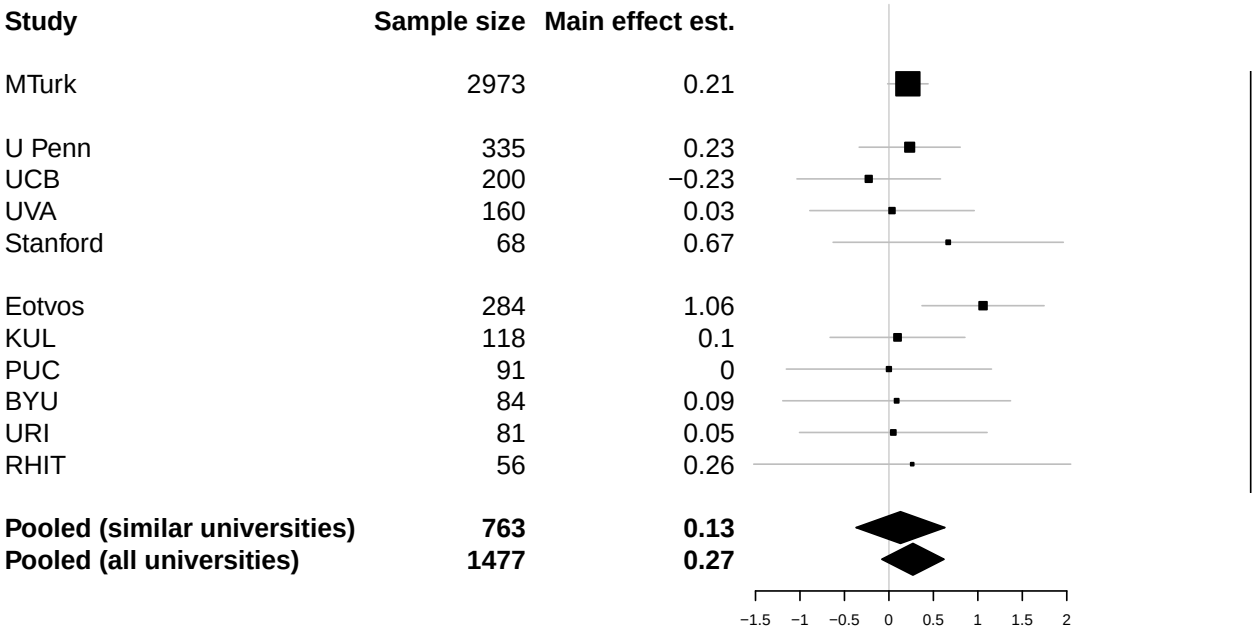


Figure 1: Forest plot for main effect estimates ordered by site type (MTurk, similar, dissimilar) and then by sample size. Point estimates and 95% CIs for each site are from ordinary least squares regression fit to that site's data. Point estimates and 95% CIs for pooled estimates are from primary and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

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

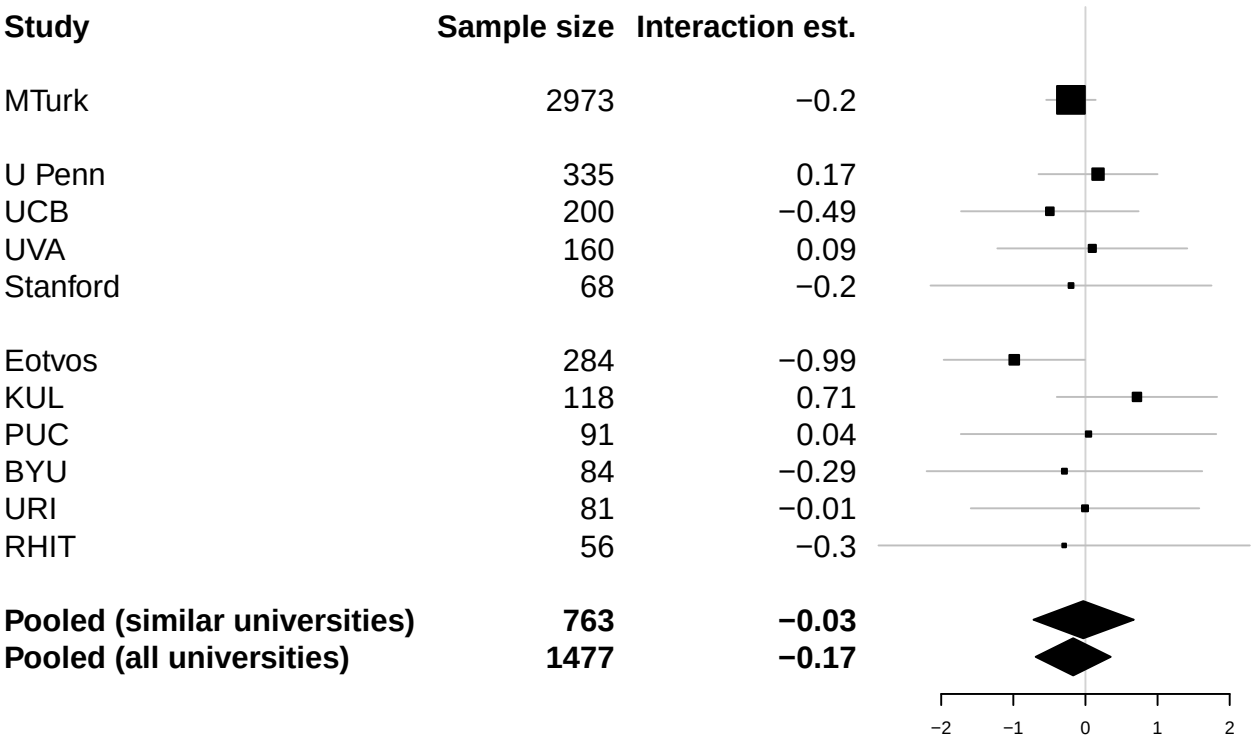


Figure 2: Forest plot for main effect estimates ordered by site type (MTurk, similar, dissimilar) and then by sample size. Point estimates and 95% CIs for each site are from ordinary least squares regression fit to that site's data. Point estimates and 95% CIs for pooled estimates are from primary and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

asssess whether the target effect differed in size between the updated protocol and the RPP protocol. To this end, we used 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; all random effects were assumed independently and identically normal[3]. This model allowed 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 with tempting fate and cognitive load). (Details of the model specification and interpretations for each coefficient of interest are provided in the preregistered protocol (CITE)). Code to perform all site-level and aggregate analyses was written by one author (MM) and audited for accuracy by other authors.

Name	Estimate	CI	pval
Magnitude of X main effect within MTurk	0.21	[-0.27, 0.7]	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

[3]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 that are provided in the Appendix.

Name	Estimate	CI	pval
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 8: 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.

Consistent with previous findings (Open Science Collaboration, 2015), the present results collected under the updated protocol in similar sites did not support the target effect (regression coefficient estimate b = -0.03 with 95% CI: [-0.72, 0.67]; p = 0.94), nor did the MTurk sample collected under the RPP protocol (b = -0.2 with 95% CI: [-0.76, 0.35]; p = 0.73). Updating the protocol did not appear to meaningfully affect the target effect (b = 0.17 with 95% CI: [-0.72, 1.06]; p = 0.75). Furthermore, results under the updated protocol also did not support the main effect of tempting fate (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, 0.7]; p = 0.67). As for the target interaction effect, 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).

Replication results in all university sites

Our first secondary analysis addressed the same questions as the primary analyses, but using all university sites rather than only similar sites. Relatively little heterogeneity was apparent across sites for both the main effect of tempting fate (estimated random intercept standard deviation = 0.23) and the target interaction (estimated random slope standard deviation = 0.38).

Name	Estimate	CI	pval
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, 0.6]	0.62
Magnitude of X-L interaction within university sites	-0.17	[-0.69, 0.35]	0.53

Name	Estimate	CI	pval
Effect of university site vs. MTurk on X-L interaction	0.03	[0.99, -0.93]	0.95

Table 9: 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.

Statistical consistency of replication results with original results

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 heterogeneity in the replications, and they can help distinguish, for example, 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 (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 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 ($P_{orig} = 0.18$) 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 in this sense with the estimated distribution of the similar replications is $P_{orig} = 0.08$; this probability increases slightly to $P_{orig} = 0.04$ when considering the distribution of all university sites.

Evaluating proposed explanations for the replication failure

We also conducted planned secondary analyses aimed at assessing the original authors' hypotheses regarding explanations for 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 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 effect of the cognitive load manipulation on the target interaction effect was nearly identical for MTurk subjects versus university subjects (-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 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 were intended as manipulation checks in the sense that an effective cognitive load manipulation would be expected to be effortful and difficult. For both perceived effort and perceived difficulty, we used subjects[4] assigned to cognitive load ($n = table(bload, \text{!is.na}(bcount.hard))[2, 2]$) to fit a linear mixed model regressing the measure 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 the former than the latter. 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; $p = 0.30$). Perceived difficulty of the task was also comparable ($b = 0.5$ with 95% CI: -0.12, 1.13; $p = 0.24$). Ultimately, these analyses do not suggest reduced effectiveness 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 (Cornell University), 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 conducted further moderation analyses to assess 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 and analyzed $n = 984$. 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 fit among subjects did not support variation by median SAT score in either the main effect of tempting fate ($b = 0.11$ for a 1-unit increase in SAT score with 95% CI: -0.27, 0.48; $p = 0.84$) or the

[4]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.

target interaction ($b = 0.01$ with 95% CI: -0.52, 0.54; $p = 0.97$).

Conclusion

For interaction: Per post hoc analyses, these results provide some evidence for statistical inconsistency of original with replications.

For main effect: Per post hoc analyses, these results weakly suggest statistical inconsistency with original, even though effect size is way smaller.

Funding

Acknowledgments

References

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.

Gerta Rücker and Martin Schumacher. Simpson's Paradox visualized: the example of the rosiglitazone meta-analysis. *BMC Medical Research Methodology*, 8(1):34, 2008.