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

Ask Charlie:

- ML3 results?
- Share results with Jane Risen?
- Anything to do before giving final draft to coauthors?

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

Introduction

- RPP results
- ML3 results

Methods

- Design

  - Endorsed vs. non-endorsed protocol manipulation
  - Details of protocol requirements for in-person sites
  - Table 1: Site characteristics, protocol details, and sample sizes (per Google Drive)
- Analysis
  - Preregistration and sample size determination (mention approx. power)
  - Primary analyses aimed to assess: (1) the target effect (interaction) using only "similar" sites to most closely approximate the author's endorsement requirements; (2) whether the target effect was different in the RPP protocol vs. the similar sites.
  - Secondary analyses aimed to assess: (1) both of the above but also including all university sites; (2) replicability of main effect; (3) statistical consistency of original with replications; (4) differences in manipulation efficacy between MTurk and universities.

Results

- Minimal descriptive stats
  - Fig 1 = 3 interaction plots (MTurk, similar, all universities)
- Results for target interaction effect
  - No evidence to support target effect in similar sites
  - Same when considering all universities
  - No evidence for differences in target effect between RPP protocol and endorsed protocol
- Results for target interaction effect
  - No evidence to support target effect in similar sites (Fig 2: forest plot for interaction)
  - Directly compare effect size to original
  - Same when considering all universities
  - No evidence for differences in target effect between RPP protocol and endorsed protocol
  - Per post hoc analyses, these results provide some evidence for statistical inconsistency of original with replications.
- Results for main effect
  - No evidence to support main effect in similar sites (Fig 3: forest plot for main effect)
  - Directly compare effect size to original
  - Same when considering all universities
  - No evidence for differences in between RPP protocol and endorsed protocol

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

Conclusion

Online Supplement: Analysis Code

Contents

Data Quality

```
# analysis dataset
setwd("~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Prepped data")
b = read.csv("prepped_data.csv")

# make dataset with only one row per site
first = b[ !duplicated(b$site), ]
# order it by site type, then by largest to smallest n
first = first[ order(first$group, -first$site.n), ]
```

```
# 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.

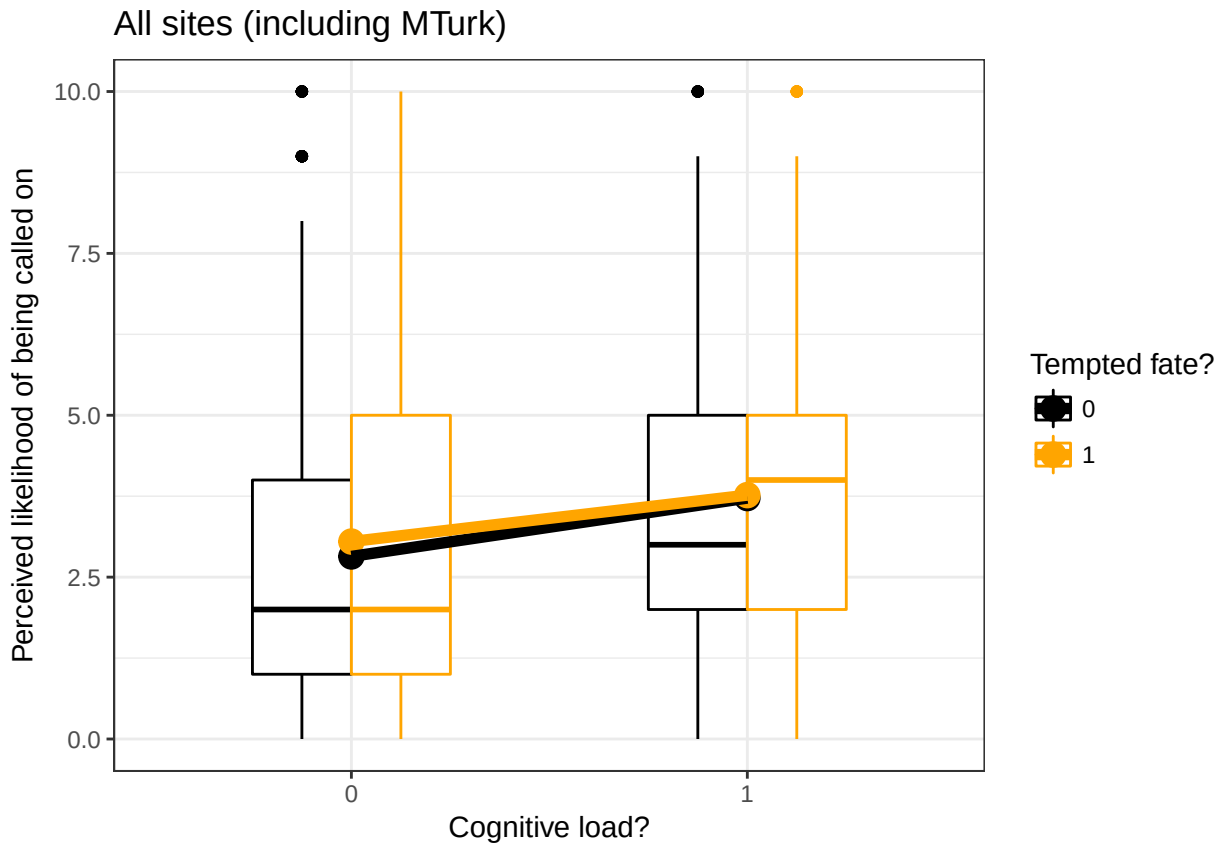
Descriptive Stats and Plots

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

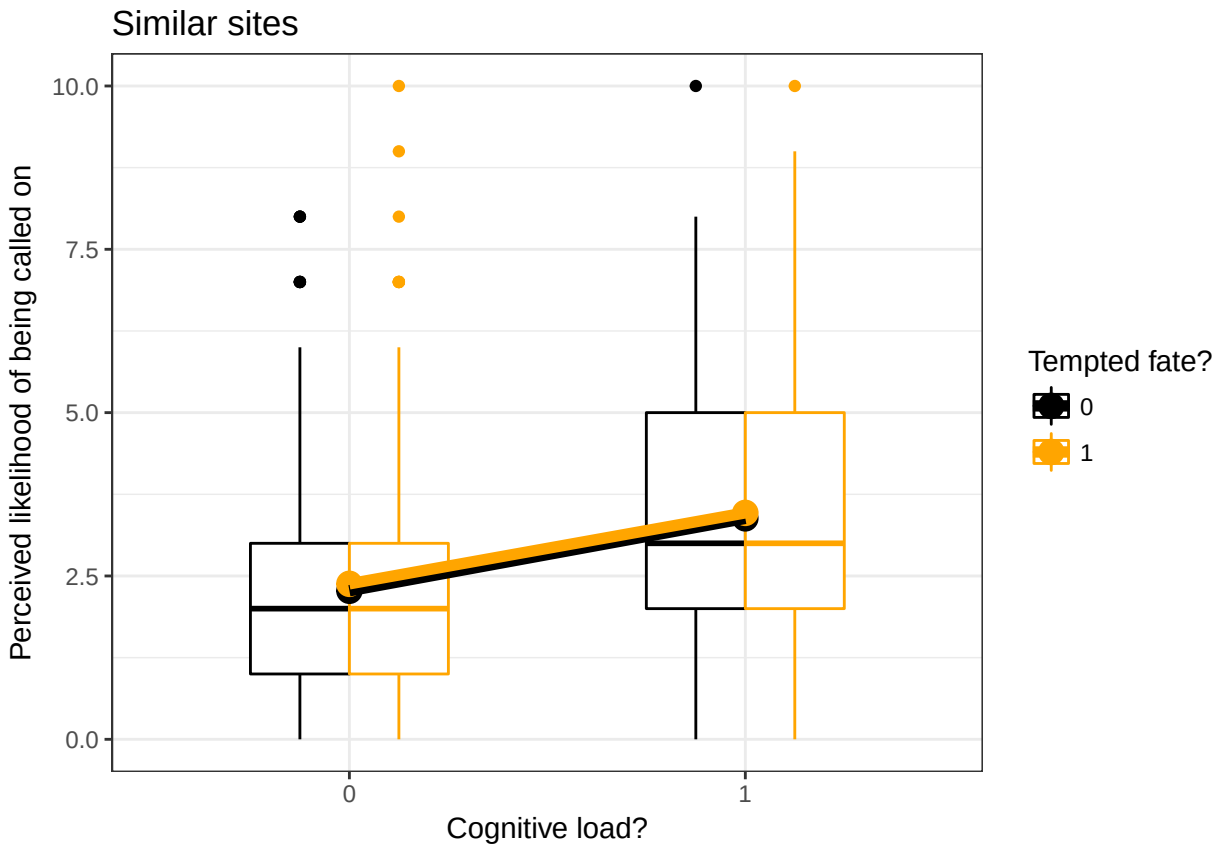
Note: These aggregated means and SDs pool across all sites within a group (similar, dissimilar, MTurk). We caution that such analyses are potentially subject to bias due to Simpson's Paradox (Rücker and Schumacher, 2008), which will be resolved in analysis models below by accounting for clustering by 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 plotting hap

  colors = c("black", "orange")
```

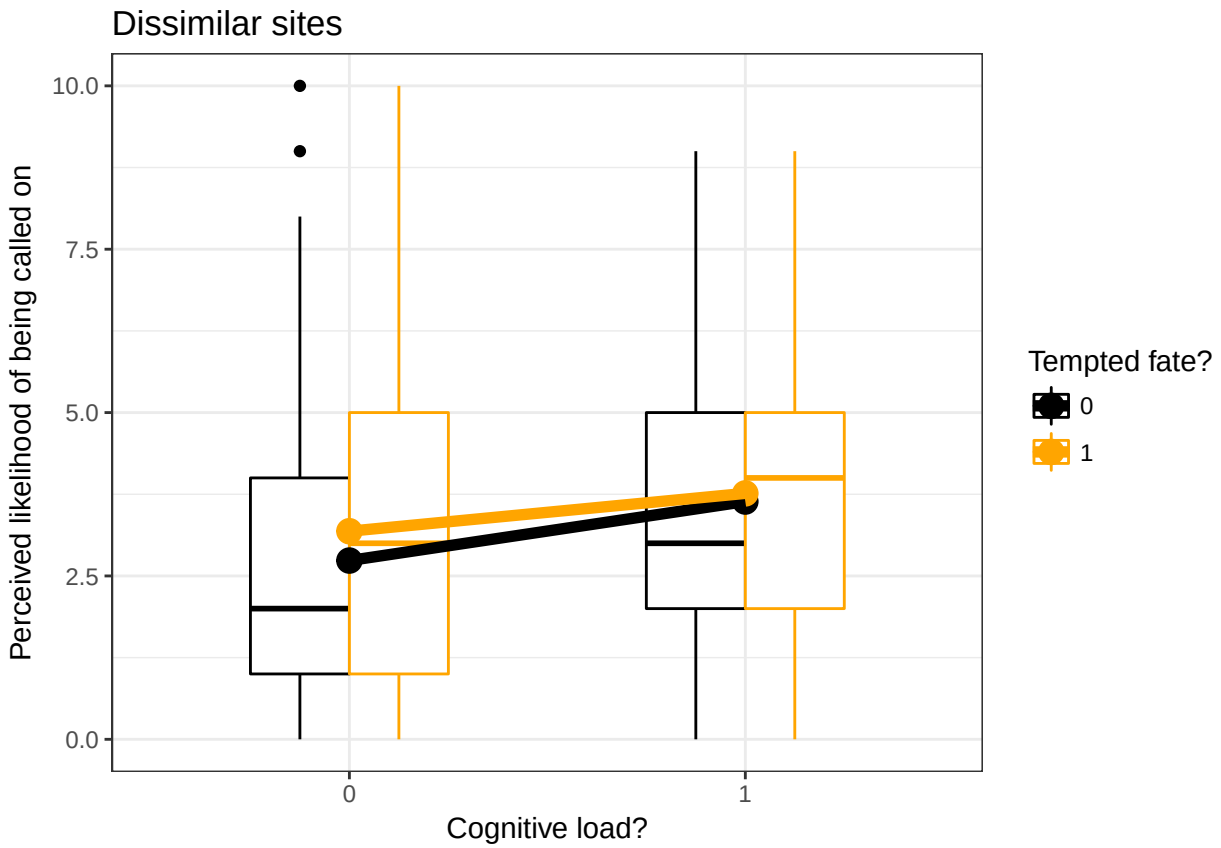
```
  plot( ggplot( dat, aes(x = as.factor(load), y = lkl, color=as.factor(tempt) ) ) + geom_boxplot(width=0.5) +
      geom_point(data = agg, aes(y = val), size=4 ) +
      geom_line(data = agg, aes(y = val, group = tempt), lwd=2 ) +
      scale_color_manual(values=colors) +
      scale_y_continuous(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



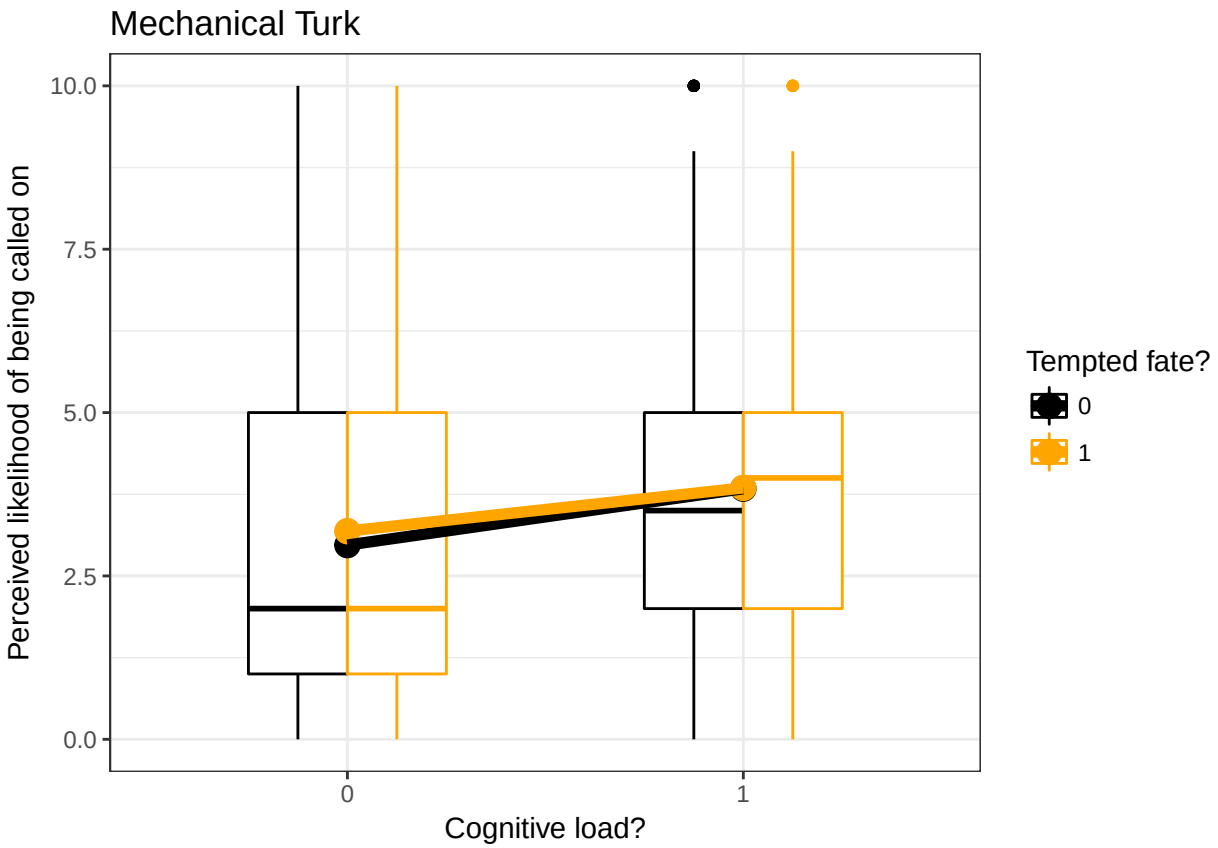

```
int_plot(b[ b$group=="b.similar", ], ggtitle = "Similar sites")
```




```
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)" )
```

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.

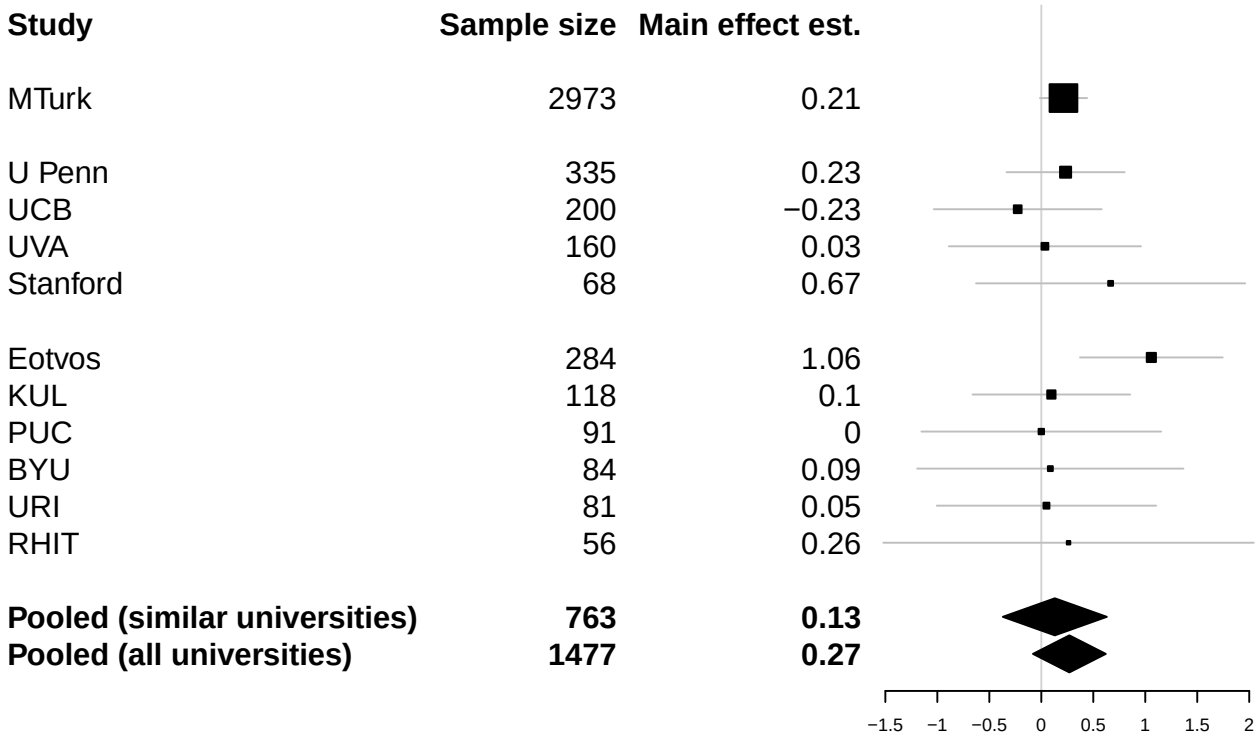
(Technical note: An alternative for the study-specific estimates would be to use estimates of random intercepts and random slopes by site from the LMM, but here we use subset analyses for a descriptive characterization that relaxes the across-site distributional assumptions of LMM.)

```
# 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 )
```

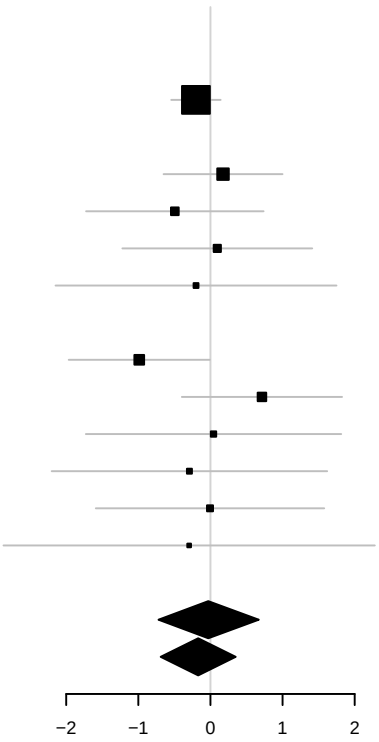
```
# make the forest plot

# 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 = 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 = c(NA, NA, pretty_spaces(first$site.main.est, use.NA = TRUE),
    NA, round(main.sim$est, 2), round(main.uni$est, 2))
l = c(NA, NA, pretty_spaces(first$site.main.lo, use.NA = TRUE),
    NA, round(main.sim$lo, 2), round(main.uni$lo, 2))
u = 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, mean = m, lower = l, upper = u,
    zero = 0, is.summary = c(TRUE, rep(FALSE, 14), TRUE, TRUE))
```




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

# build text 'columns' of forest plot NAs are for making
# spaces
tabletext = 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)))

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

```
# 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 interaction")

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)

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

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 interaction

# 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)

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

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.

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 = 2.93 - 1.9
var.mean0 = 1.42^2/30
var.mean1 = 2.16^2/30
vyi.orig = var.mean0 + var.mean1

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

```
## [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 = p_orig(orig.y = yi.orig, orig.vy = vyi.orig, 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, orig.vy = vyi.orig, 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, orig.vy = vyi.orig, 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 = (5.27 - 2.7) - (2.93 - 1.9)
vyi.orig = (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
(yi.orig/sqrt(vyi.orig))^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(orig.y = yi.orig, orig.vy = vyi.orig, 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] 0.08155653
```

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, orig.vy = vyi.orig, 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, orig.vy = vyi.orig, 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.

```
# effect sizes of original
yi.orig = 2.93 - 1.9
vyi.orig = (1.42^2 + 2.16^2)/2  # within-study variance of the difference
```

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(orig.y = yi.orig, orig.vy = vyi.orig, 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] 0.6809201
```

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, orig.vy = vyi.orig, 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] 0.6238139
```

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

10 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)

```
# interaction is the 'difference in differences'
yi.orig = (5.27 - 2.7) - (2.93 - 1.9)
vyi.orig = (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
```

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 = p_orig(orig.y = yi.orig, orig.vy = vyi.orig, 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, orig.vy = vyi.orig, 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, orig.vy = vyi.orig, 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 = 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 = confint(m, 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
(m3 = 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
```

```
CI3 = confint(m3, method = "Wald")

(m4 = 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
CI4 = confint(m4, 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

```
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
```

```
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.

References

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

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