ML5 Risen & Gilovich Data Preparation

Contact: Maya Mathur (mmathur@stanford.edu)

June 4, 2017

Site-Level Data Preparation

Overview: A central script is called to prep each site's data automatically (merging various columns, producing standardized names, and excluding subjects per the a priori attention criterion) while outputting results of sanity checks and producing a within-site interaction plot. The script writes separate files with each site's prepped data. Lastly, all sites' prepped data are stitched into a single analysis dataset.

Plots show standard boxplots (quartiles) with lines overlaying group means.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Brigham Young

First manually add the `had.read` and `load` variables:

```
source("~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/3. Analysis/R c

setwd("~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Raw data/
d = read.csv("raw_byu.csv", header = TRUE)

# remove additional header rows
d = d[-c(1:2), ]

library(plyr)
library(dplyr)

# variable names here are sometimes exactly the same rename
# them to avoid problems
names(d)[18:(length(names(d)) - 1)] = c("lkl1", "imp1", "bad1",
    "lkl2", "endnum1", "eff.split1", "count.hard1", "count.eff1",
    "imp2", "bad2", "lkl3", "endnum2", "eff.split2", "count.hard2",
```

```
    "count.eff2", "imp3", "bad3", "lkl4", "imp4", "bad4")

# make had.read variable
d$had.read = NA
d$had.read[!is.na(as.numeric(as.character(d$lkl1))) | !is.na(as.numeric(as.character(d$lkl3)))] = 0

d$had.read[!is.na(as.numeric(as.character(d$lkl2))) | !is.na(as.numeric(as.character(d$lkl4)))] = 1

# merge end-number columns warning about 'NAs introduced by
# coercion', but is correct
d$end.num = coalesce(as.numeric(as.character(d$endnum1)), as.numeric(as.character(d$endnum2)))

# make load variable based on end-number column
d$load = 0
d$load[!is.na(as.numeric(as.character(d$end.num)))] = 1


# merge effort-split columns warning about 'NAs introduced by
# coercion', but is correct
d$eff.split = coalesce(as.numeric(as.character(d$eff.split1)),
    as.numeric(as.character(d$eff.split2)))

# merge badness columns warning about 'NAs introduced by
# coercion', but is correct
d$badness = coalesce(as.numeric(as.character(d$bad1)), as.numeric(as.character(d$bad2)),
    as.numeric(as.character(d$bad3)), as.numeric(as.character(d$bad4)))

# merge importance columns warning about 'NAs introduced by
# coercion', but is correct
d$importance = coalesce(as.numeric(as.character(d$imp1)), as.numeric(as.character(d$imp2)),
    as.numeric(as.character(d$imp3)), as.numeric(as.character(d$imp4)))

# merge counting effort columns warning about 'NAs introduced
# by coercion', but is correct
d$count.eff = coalesce(as.numeric(as.character(d$count.eff1)),
    as.numeric(as.character(d$count.eff2)))

# merge counting effort columns warning about 'NAs introduced
# by coercion', but is correct
d$count.hard = coalesce(as.numeric(as.character(d$count.hard1)),
    as.numeric(as.character(d$count.hard2)))

write.csv(d, "manualprep_byu.csv")
```

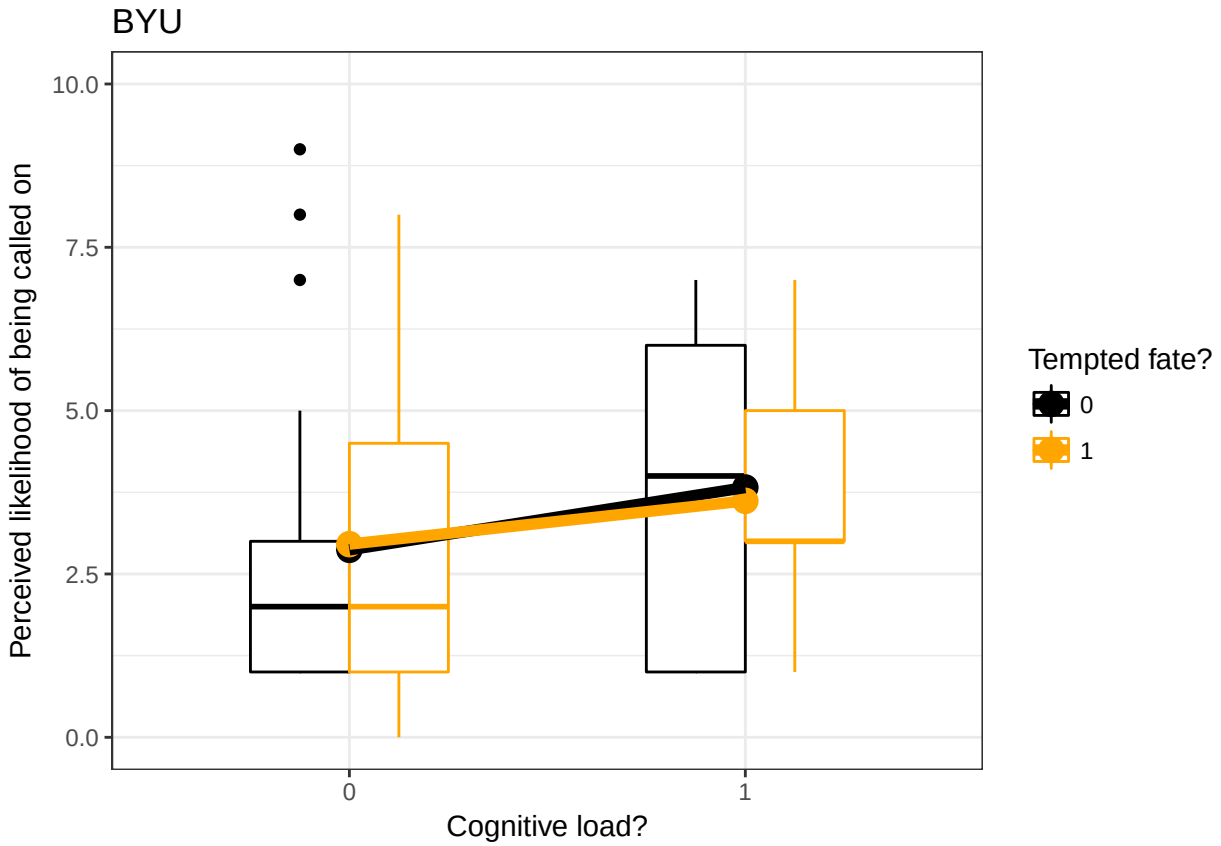
Automatic data prep:

```
start.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Ra
end.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Prep

source("~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/3. Analysis/R co
prep_site_data(start.path = start.path, end.path = end.path,
    lkl.names = c("lkl1", "lkl2", "lkl3", "lkl4"), had.read.name = "had.read",
    load.name = "load", end.num.name = "end.num", eff.split.name = "eff.split",
```

```
    count.eff.name = "count.eff", count.hard.name = "count.hard",
    badness.name = "badness", importance.name = "importance",
    .site.name = "BYU", .group = "c.dissimilar", .n.extra.header.rows = 0)
```

```
##
##
## No extra header rows to delete.
##
## Rows in raw data = 90
##
## Head of skinny dataset before exclusions:
##   id .site.name       .group had.read load lkl eff.split count.eff
## 1  1        BYU c.dissimilar        0    1   6         2         5
## 2  2        BYU c.dissimilar        1    0   3        NA        NA
## 3  3        BYU c.dissimilar        0    0   6        NA        NA
## 4  4        BYU c.dissimilar        1    0   1        NA        NA
## 5  5        BYU c.dissimilar        1    1   1         1        10
## 6  6        BYU c.dissimilar        0    1   1         1         9
##   count.hard badness importance end.num
## 1          5       9          8     549
## 2         NA       4          2      NA
## 3         NA       4          4      NA
## 4         NA       5          8      NA
## 5         10      10         10     548
## 6          8       4          8     552
##
##
## Subjects with missing had.read, load, or lkl:
##
## Subjects with load==1 but missing eff.split, count.eff, or count.hard:
##
## Bad subjects (failed to follow instructions): 11 17 19 35 81 85
##
## Final n = 84
##
## MARGINAL MEANS AND SDs FOR ANALYSIS AUDIT
##                     Overall
##   n                   84
##   load (mean (sd))  0.45 (0.50)
##   tempt (mean (sd)) 0.52 (0.50)
##   lkl (mean (sd))   3.27 (2.18)
```



Eotvos Lorand

First manually exclude 7 subjects who may have completed the experiment twice (note: these 7 are on top of any "bad subjects" excluded by the prep script):

```
source("~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/3. Analysis/R c

setwd("~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Raw data/
d = read.csv("raw_eotvos.csv")

d = d[!d$ResponseId %in% c("R_1H76w6p0V7mQW3T", "R_12mU6WnhCkSe5LH",
    "R_3ReInhemqnhITwd", "R_2tqY0auQ9EUmWCf", "R_Z4tlLy5HCyAXhRf",
    "R_28BCOFVHOcgVquF", "R_1LwH5vC3QBq4uFG"), ]

write.csv(d, "manualprep_eotvos.csv")
```

Automatic data prep:

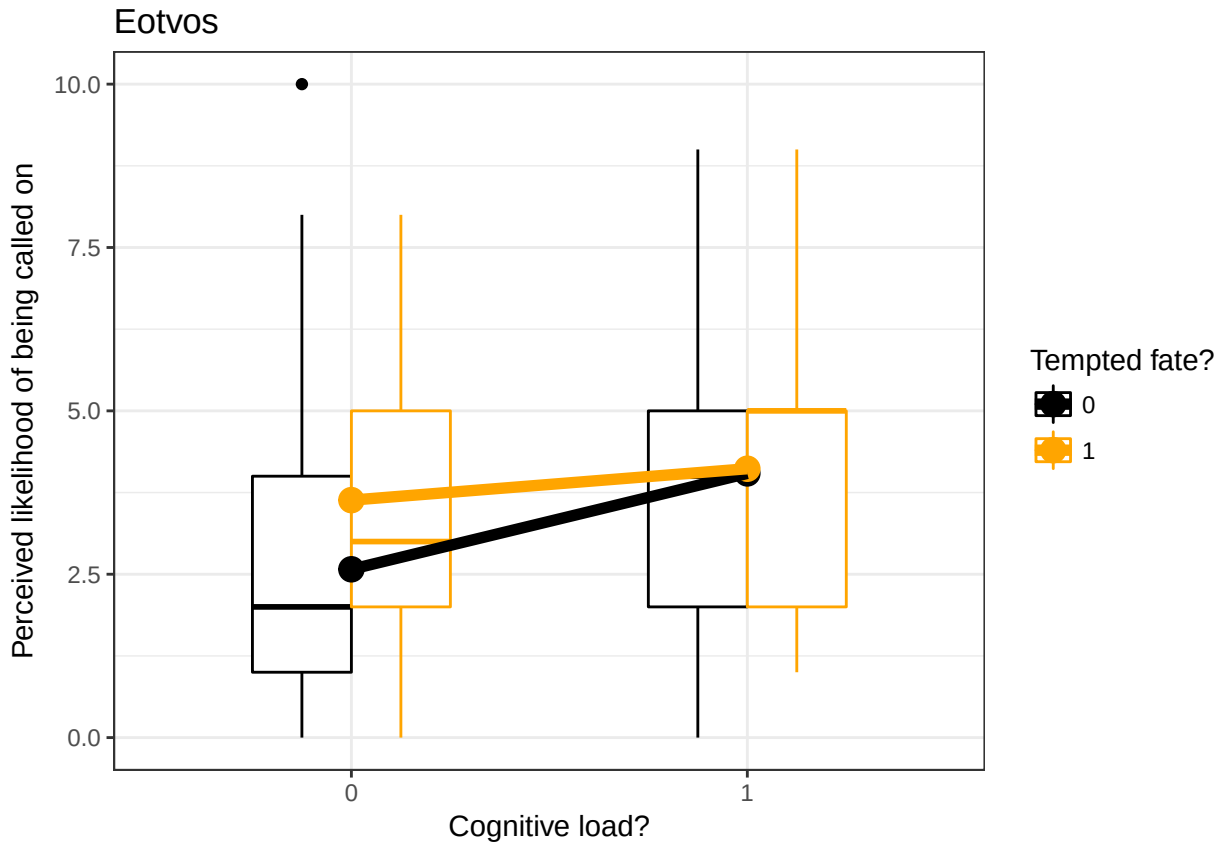
```
start.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Ra
end.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Prep

prep_site_data(start.path = start.path, end.path = end.path,
    lkl.names = c("L1.R1.scenario_1", "L1.R0.scenario_1", "L0.R1.text_1",
        "L0.R0.text_1"), had.read.name = "had.read", load.name = "load",
    end.num.name = "Q28", eff.split.name = "Q29_1", count.eff.name = "Q22_1",
```

```
    count.hard.name = "Q21_1", badness.name = "Q14_1", importance.name = "Q26_1",
    .site.name = "Eotvos", .group = "c.dissimilar", .n.extra.header.rows = 2)
```

```
##
##
## Extra header rows to delete (first 3 cols):
##   X                StartDate                EndDate
## 1 1               Start Date               End Date
## 2 2 {"ImportId":"startDate"} {"ImportId":"endDate"}
##
##
## First row of real data:
##   X           StartDate             EndDate Status      IPAddress Progress
## 3 3 2017-02-28 10:26:32 2017-02-28 10:27:56      0 157.181.60.140      100
##   Duration..in.seconds. Finished        RecordedDate        ResponseId
## 3                    83        1 2017-02-28 10:27:57 R_3FOLtIrDKwpIDjL
##   RecipientLastName RecipientFirstName RecipientEmail ExternalReference
## 3
##   LocationLatitude LocationLongitude DistributionChannel Q15 Q16
## 3             47.5   19.083297729492           anonymous   1   1
##   L1.R1.scenario_1 L1.R0.scenario_1 Q28 Q29_1 Q21_1 Q22_1 L0.R1.text_1
## 3                                                                    1
##   L0.R0.text_1 Q26_1 Q14_1 mTurkCode load had.read
## 3                7    10    423397    0        1
##
## Rows in raw data = 291
##
## Head of skinny dataset before exclusions:
##   id .site.name       .group had.read load lkl eff.split count.eff
## 1  1     Eotvos c.dissimilar        1    0   1        NA        NA
## 2  2     Eotvos c.dissimilar        0    0   3        NA        NA
## 3  3     Eotvos c.dissimilar        1    0   5        NA        NA
## 4  4     Eotvos c.dissimilar        1    0   2        NA        NA
## 5  5     Eotvos c.dissimilar        0    0   4        NA        NA
## 6  6     Eotvos c.dissimilar        1    1   3         4         7
##   count.hard badness importance end.num
## 1         NA      10          7      NA
## 2         NA       7          8      NA
## 3         NA      10          3      NA
## 4         NA       3          5      NA
## 5         NA       5          7      NA
## 6          6       9          9     534
##
##
## Subjects with missing had.read, load, or lkl:
##
## Subjects with load==1 but missing eff.split, count.eff, or count.hard: 170
##
## Bad subjects (failed to follow instructions): 11 28 84 140 169 193 266
##
## Final n = 284
##
## MARGINAL MEANS AND SDs FOR ANALYSIS AUDIT
```

```
##                      Overall
##   n                   284
##   load (mean (sd))  0.49 (0.50)
##   tempt (mean (sd)) 0.50 (0.50)
##   lkl (mean (sd))   3.58 (2.17)
```



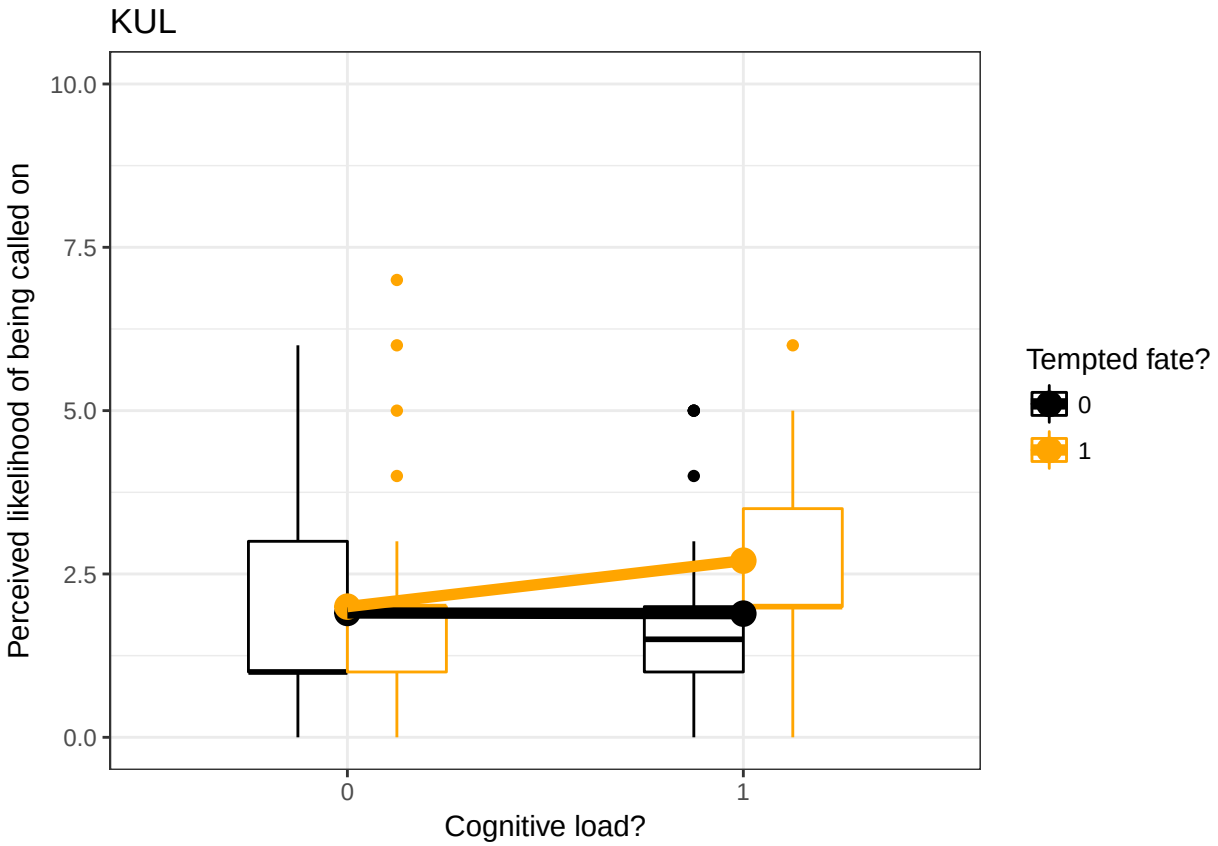
KU Leuven

```
start.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Ra
end.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Prep

prep_site_data(start.path = start.path, end.path = end.path,
    lkl.names = c("L1.R1.scenario_1", "L1.R0.scenario_1", "L0.R1.text_1",
        "L0.R0.text_1"), had.read.name = "had.read", load.name = "load",
    end.num.name = "Q28", eff.split.name = "Q29_1", count.eff.name = "Q22_1",
    count.hard.name = "Q21_1", badness.name = "Q14_1", importance.name = "Q26_1",
    .site.name = "KUL", .group = "c.dissimilar", .n.extra.header.rows = 1)
```

```
##
##
## Extra header rows to delete (first 3 cols):
##           V1          V2   V3
## 1 ResponseID ResponseSet Name
##
```

```
##
## First row of real data:
##                  V1                   V2        V3 V4 V5             V6 V7
## 2 R_1FPzjjMqHZwdcbk Default Response Set Anonymous       134.58.253.57  0
##              V8            V9 V10 load had.read text L1.R1.scenario_1
## 2 2/14/17 10:06 2/14/17 10:07   1    0        1
##   L1.R0.scenario_1 Q28 Q29_1 Q21_1 Q22_1 L0.R1.text_1 L0.R0.text_1 Q26_1
## 2                                                 2                 5
##   Q14_1 text.thanks   LocationLatitude  LocationLongitude LocationAccuracy
## 2     8           1 50,879,592,895,508 47,008,972,167,969               -1
##
## Rows in raw data = 127
##
## Head of skinny dataset before exclusions:
##   id .site.name       .group had.read load lkl eff.split count.eff
## 1  1        KUL c.dissimilar        1    0   2        NA        NA
## 2  2        KUL c.dissimilar        1    1   6         2         4
## 3  3        KUL c.dissimilar        0    1   2         4         9
## 4  4        KUL c.dissimilar        0    0   7        NA        NA
## 5  5        KUL c.dissimilar        1    1   2         3         9
## 6  6        KUL c.dissimilar        0    1   2         4         5
##   count.hard badness importance end.num
## 1         NA       8          5      NA
## 2          3       3          2     561
## 3          7       9          7     525
## 4         NA       8          9      NA
## 5          8       7          6     343
## 6          4       9          8     504
##
##
## Subjects with missing had.read, load, or lkl:
##
## Subjects with load==1 but missing eff.split, count.eff, or count.hard:
##
## Bad subjects (failed to follow instructions): 2 30 34 65 71 85 89 102 106
##
## Final n = 118
##
## MARGINAL MEANS AND SDs FOR ANALYSIS AUDIT
##                      Overall
##   n                   118
##   load (mean (sd))  0.47 (0.50)
##   tempt (mean (sd)) 0.50 (0.50)
##   lkl (mean (sd))   2.11 (1.53)
```



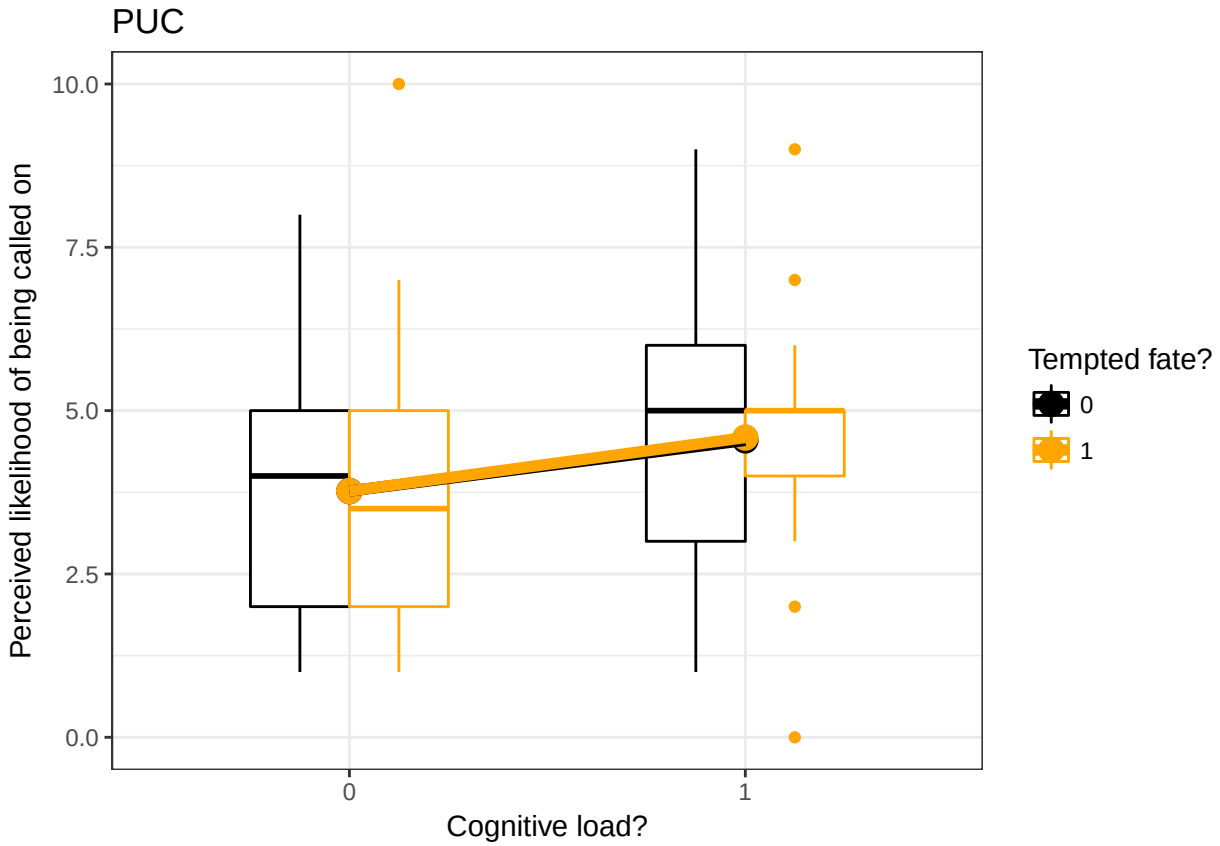
PUC Rio

```
start.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Ra
end.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Prep

prep_site_data(start.path = start.path, end.path = end.path,
    lkl.names = c("L1.R1.scenario_1", "L1.R0.scenario_1", "L0.R1.text_1",
        "L0.R0.text_1"), had.read.name = "had.read", load.name = "load",
    end.num.name = "Q28", eff.split.name = "Q29_1", count.eff.name = "Q22_1",
    count.hard.name = "Q21_1", badness.name = "Q14_1", importance.name = "Q26_1",
    .site.name = "PUC", .group = "c.dissimilar", .n.extra.header.rows = 1)
```

```
##
##
## Extra header rows to delete (first 3 cols):
##           V1          V2   V3
## 1 ResponseID ResponseSet Name
##
##
## First row of real data:
##                  V1                   V2        V3 V4 V5             V6
## 2 R_2tnZTPfaEpn2cC2 Default Response Set Anonymous       179.233.112.179
##   V7                  V8                  V9 V10 mTurkCode load had.read
## 2  0 2017-02-15 16:37:42 2017-02-15 16:37:49   0    423812    0        0
```

```
##   Q24 Q13 text L1.R1.scenario_1 L1.R0.scenario_1 Q28 Q29_1 Q21_1 Q22_1
## 2   1
##   L0.R1.text_1 L0.R0.text_1 Q26_1 Q14_1 text.thanks LocationLatitude
## 2
##   LocationLongitude LocationAccuracy  X
## 2                                -1 NA
##
## Rows in raw data = 106
##
## Head of skinny dataset before exclusions:
##   id .site.name       .group had.read load lkl eff.split count.eff
## 1  1        PUC c.dissimilar        0    0  NA        NA        NA
## 2  2        PUC c.dissimilar        0    0  NA        NA        NA
## 3  3        PUC c.dissimilar        0    0   6        NA        NA
## 4  4        PUC c.dissimilar        1    0   4        NA        NA
## 5  5        PUC c.dissimilar        1    1   2         3         4
## 6  6        PUC c.dissimilar        0    1   5         2         6
##   count.hard badness importance end.num
## 1         NA      NA         NA      NA
## 2         NA      NA         NA      NA
## 3         NA       9          9      NA
## 4         NA       6          6      NA
## 5          4       6          5     555
## 6          5       5          5     537
##
##
## Subjects with missing had.read, load, or lkl: 1 2
##
## Subjects with load==1 but missing eff.split, count.eff, or count.hard: 15 106
##
## Bad subjects (failed to follow instructions): 13 14 25 34 35 41 52 54 58 69 85 89 102
##
## Final n = 91
##
## MARGINAL MEANS AND SDs FOR ANALYSIS AUDIT
##                     Overall
##   n                   91
##   load (mean (sd))  0.43 (0.50)
##   tempt (mean (sd)) 0.47 (0.50)
##   lkl (mean (sd))   4.11 (2.09)
```



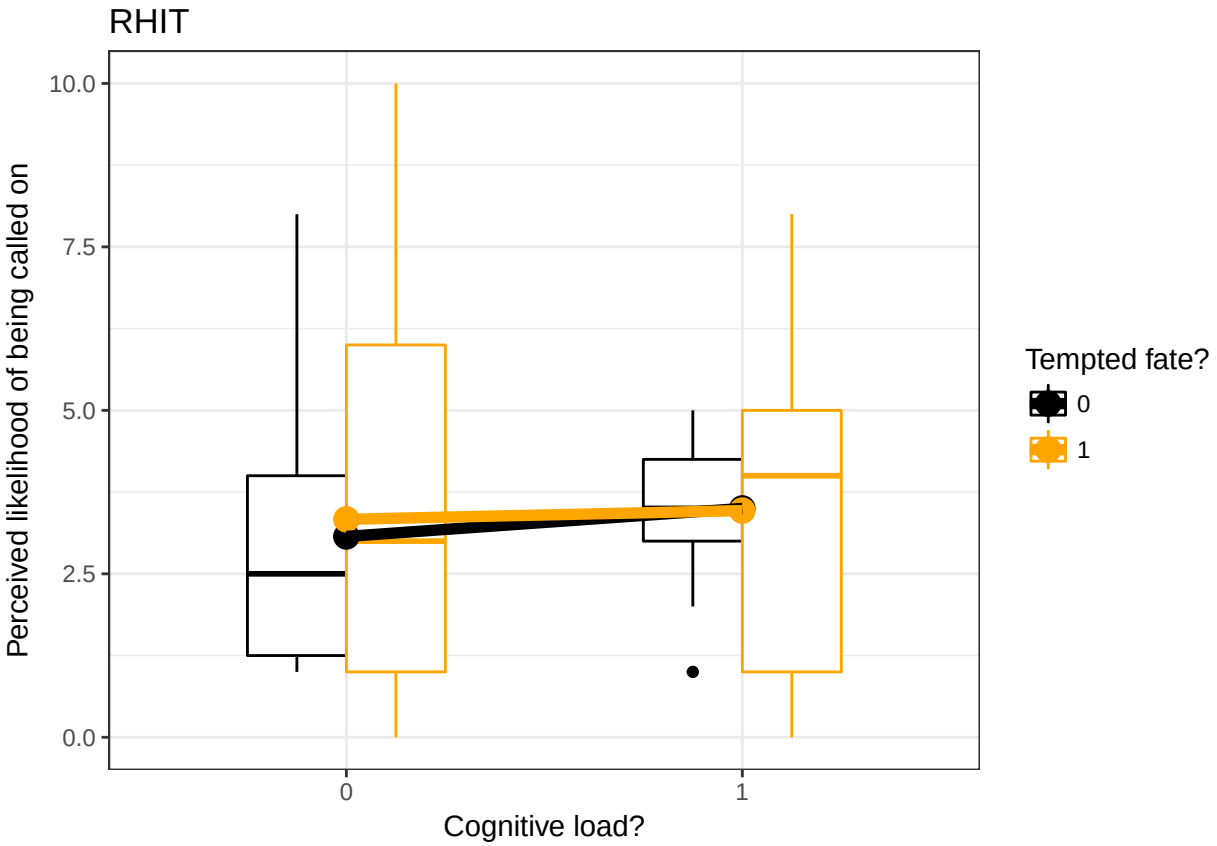
Rose-Hulman IT

```
start.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Ra
end.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Prep

d = prep_site_data(start.path = start.path, end.path = end.path,
    lkl.names = c("likelihood_1", "likelihood_1.1", "likelihood_1.2",
        "likelihood_1.3"), had.read.name = "had.read", load.name = "load",
    end.num.name = "end.num", eff.split.name = "effort.split_1",
    count.eff.name = "effort.count_1", count.hard.name = "difficulty_1",
    badness.name = "negativity_1", importance.name = "academic.pressure_1",
    .site.name = "RHIT", .group = "c.dissimilar", .n.extra.header.rows = 2)
```

```
##
##
## Extra header rows to delete (first 3 cols):
##                   StartDate                 EndDate                  Status
## 1                Start Date                End Date           Response Type
## 2 {"ImportId":"startDate"} {"ImportId":"endDate"} {"ImportId":"status"}
##
##
## First row of real data:
##              StartDate             EndDate     Status       IPAddress
## 3 2017-01-05 18:49:40 2017-01-05 18:51:42 IP Address 137.112.236.167
```

```
##   Progress Duration..in.seconds. Finished        RecordedDate
## 3      100                   121     True 2017-01-05 18:51:43
##          ResponseId RecipientLastName RecipientFirstName RecipientEmail
## 3 R_1r64G1b4VqU0tqw
##   ExternalReference LocationLatitude LocationLongitude DistributionChannel
## 3                   39.466201782227  -87.314796447754           anonymous
##   likelihood_1 likelihood_1.1 end.num effort.split_1 difficulty_1
## 3                           8     547              6            1
##   effort.count_1 likelihood_1.2 likelihood_1.3 academic.pressure_1
## 3              1                                                 9
##   negativity_1 mTurkCode load had.read
## 3            9    423608    1        0
##
## Rows in raw data = 58
##
## Head of skinny dataset before exclusions:
##   id .site.name       .group had.read load lkl eff.split count.eff
## 1  1       RHIT c.dissimilar        0    1   8         6         1
## 2  2       RHIT c.dissimilar        0    0  10        NA        NA
## 3  3       RHIT c.dissimilar        1    0   1        NA        NA
## 4  4       RHIT c.dissimilar        1    1   3         3         6
## 5  5       RHIT c.dissimilar        0    0   6        NA        NA
## 6  6       RHIT c.dissimilar        0    1   3         3         5
##   count.hard badness importance end.num
## 1          1       9          9     547
## 2         NA      10         10      NA
## 3         NA       7          8      NA
## 4          4       8          6     543
## 5         NA       6          3      NA
## 6          6       3          7     540
##
##
## Subjects with missing had.read, load, or lkl:
##
## Subjects with load==1 but missing eff.split, count.eff, or count.hard: 15
##
## Bad subjects (failed to follow instructions): 32 33
##
## Final n = 56
##
## MARGINAL MEANS AND SDs FOR ANALYSIS AUDIT
##                     Overall
##   n                   56
##   load (mean (sd))  0.48 (0.50)
##   tempt (mean (sd)) 0.54 (0.50)
##   lkl (mean (sd))   3.34 (2.33)
```



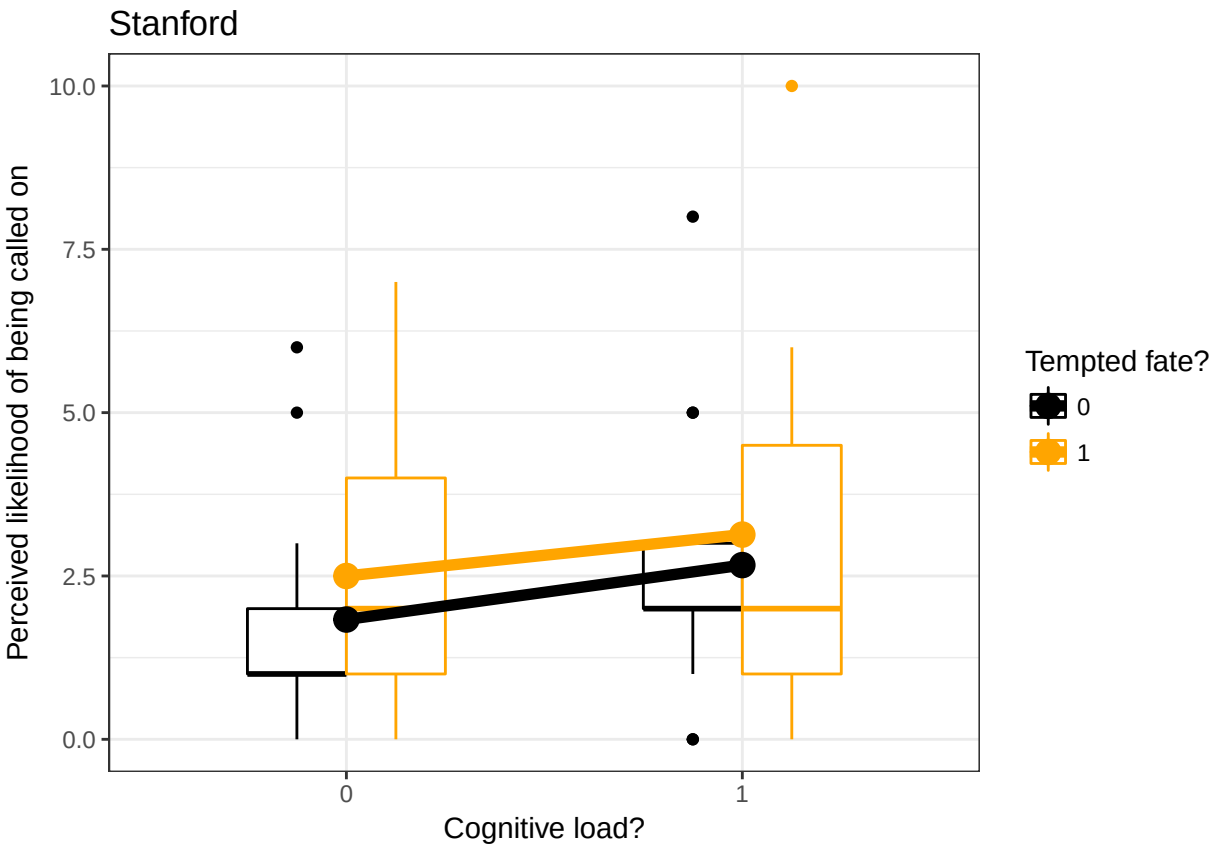
Stanford

```
start.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Ra
end.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Prep

prep_site_data(start.path = start.path, end.path = end.path,
    lkl.names = c("likelihood_1", "likelihood_1.1", "likelihood_1.2",
        "likelihood_1.3"), had.read.name = "had.read", load.name = "load",
    end.num.name = "end.num", eff.split.name = "effort.split_1",
    count.eff.name = "effort.count_1", count.hard.name = "difficulty_1",
    badness.name = "negativity_1", importance.name = "academic.pressure_1",
    .site.name = "Stanford", .group = "b.similar", .n.extra.header.rows = 2)
```

```
##
##
## Extra header rows to delete (first 3 cols):
##                   StartDate                 EndDate                  Status
## 1                Start Date                End Date           Response Type
## 2 {"ImportId":"startDate"} {"ImportId":"endDate"} {"ImportId":"status"}
##
##
## First row of real data:
##              StartDate             EndDate     Status     IPAddress
## 3 2016-11-28 09:34:37 2016-11-28 09:36:33 IP Address 68.65.174.224
```

```
##   Progress Duration..in.seconds. Finished        RecordedDate
## 3      100                   115     True 2016-11-28 09:36:34
##          ResponseId RecipientLastName RecipientFirstName RecipientEmail
## 3 R_3fGZ45IO9Zkbbdm
##   ExternalReference LocationLatitude LocationLongitude DistributionChannel
## 3                   37.41780090332  -122.17199707031           anonymous
##   likelihood_1 likelihood_1.1 end.num effort.split_1 difficulty_1
## 3
##   effort.count_1 likelihood_1.2 likelihood_1.3 academic.pressure_1
## 3                                           1                   2
##   negativity_1 mTurkCode load had.read
## 3           10    423884    0        0
##
## Rows in raw data = 74
##
## Head of skinny dataset before exclusions:
##   id .site.name    .group had.read load lkl eff.split count.eff count.hard
## 1  1   Stanford b.similar        0    0   1        NA        NA         NA
## 2  2   Stanford b.similar        1    1   5         4         6          4
## 3  3   Stanford b.similar        0    0   7        NA        NA         NA
## 4  4   Stanford b.similar        1    0   1        NA        NA         NA
## 5  5   Stanford b.similar        0    0   3        NA        NA         NA
## 6  6   Stanford b.similar        0    1   2         4         3          1
##   badness importance end.num
## 1      10          2      NA
## 2       7          7     537
## 3       9         10      NA
## 4       8          4      NA
## 5       7          7      NA
## 6       6          3     525
##
##
## Subjects with missing had.read, load, or lkl: 10 21 31 37 38
##
## Subjects with load==1 but missing eff.split, count.eff, or count.hard: 10 21 22 31 37 38
##
## Bad subjects (failed to follow instructions): 18
##
## Final n = 68
##
## MARGINAL MEANS AND SDs FOR ANALYSIS AUDIT
##                     Overall
##   n                   68
##   load (mean (sd))  0.44 (0.50)
##   tempt (mean (sd)) 0.51 (0.50)
##   lkl (mean (sd))   2.50 (2.00)
```



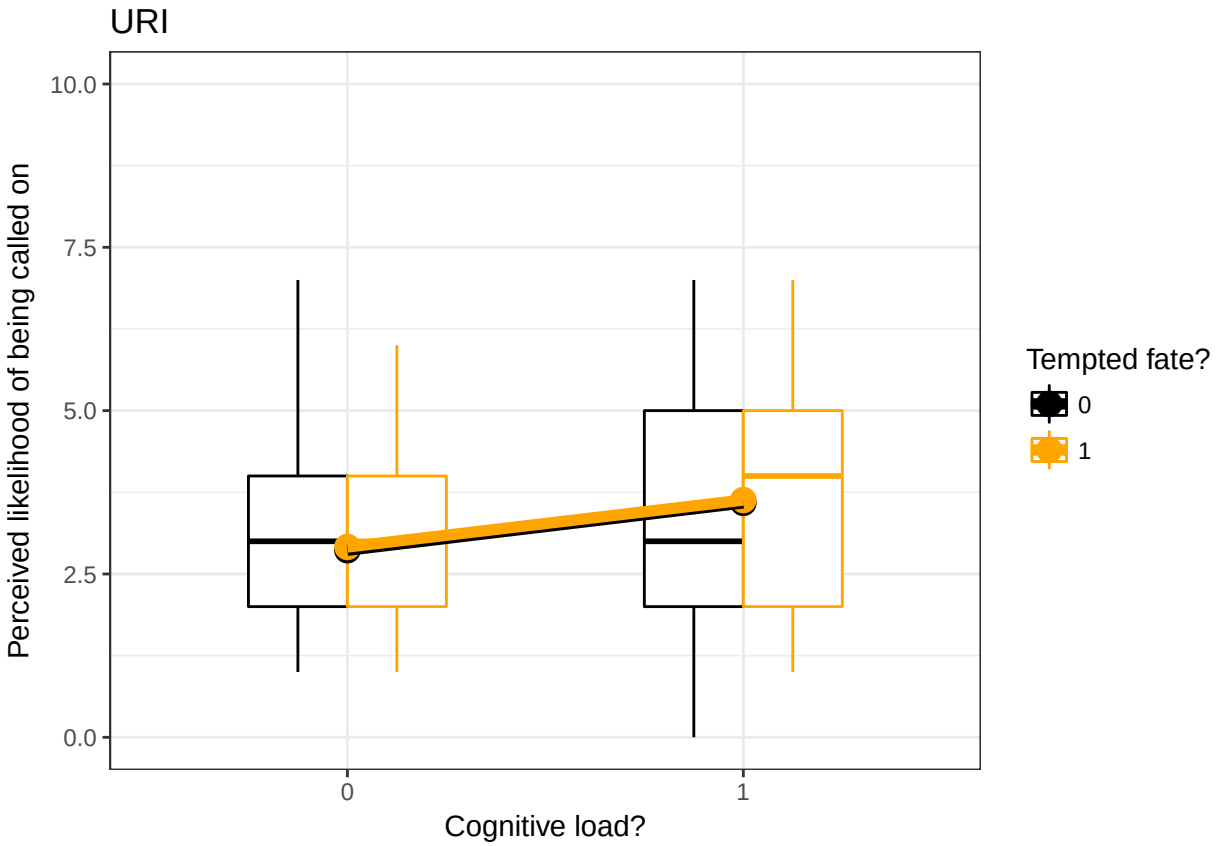
University of Rhode Island

```
start.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Ra
end.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Prep

prep_site_data(start.path = start.path, end.path = end.path,
    lkl.names = c("L1.R1.scenario_1", "L1.R0.scenario_1", "L0.R1.text_1",
        "L0.R0.text_1"), had.read.name = "had.read", load.name = "load",
    end.num.name = "Q28", eff.split.name = "Q29_1", count.eff.name = "Q22_1",
    count.hard.name = "Q21_1", badness.name = "Q14_1", importance.name = "Q26_1",
    .site.name = "URI", .group = "c.dissimilar", .n.extra.header.rows = 2)
```

```
##
##
## Extra header rows to delete (first 3 cols):
##                   StartDate                 EndDate                Progress
## 1                Start Date                End Date                Progress
## 2 {"ImportId":"startDate"} {"ImportId":"endDate"} {"ImportId":"progress"}
##
##
## First row of real data:
##       StartDate       EndDate Progress Duration..in.seconds. Finished
## 3 2/7/2017 7:13 2/7/2017 7:22      100                   568     TRUE
##    RecordedDate        ResponseId RecipientLastName RecipientFirstName
```

```
## 3 2/7/2017 7:22 R_9Rk3p8c1pP4sXUl
##   RecipientEmail ExternalReference LocationLatitude LocationLongitude
## 3                                     41.47729492      -71.52079773
##   DistributionChannel L1.R1.scenario_1 L1.R0.scenario_1 Q28 Q29_1 Q21_1
## 3           anonymous                                 2 534     4     9
##   Q22_1 L0.R1.text_1 L0.R0.text_1 Q26_1 Q14_1 mTurkCode load had.read
## 3     9                          8     6    423114    1        0
##
## Rows in raw data = 90
##
## Head of skinny dataset before exclusions:
##   id .site.name       .group had.read load lkl eff.split count.eff
## 1  1        URI c.dissimilar        0    1   2         4         9
## 2  2        URI c.dissimilar        0    1   3         0         6
## 3  3        URI c.dissimilar        1    0   1        NA        NA
## 4  4        URI c.dissimilar        1    1   5         2         2
## 5  5        URI c.dissimilar        1    0   1        NA        NA
## 6  6        URI c.dissimilar        0    1   4         3         7
##   count.hard badness importance end.num
## 1          9       6          8     534
## 2          3      10          2     559
## 3         NA       9         10      NA
## 4          3       6          5     534
## 5         NA       9          7      NA
## 6          9       8          7     443
##
##
## Subjects with missing had.read, load, or lkl:
##
## Subjects with load==1 but missing eff.split, count.eff, or count.hard: 30 33 44 49
##
## Bad subjects (failed to follow instructions): 2 9 14 17 44 48 49 60 61
##
## Final n = 81
##
## MARGINAL MEANS AND SDs FOR ANALYSIS AUDIT
##                      Overall
##   n                    81
##   load (mean (sd))  0.44 (0.50)
##   tempt (mean (sd)) 0.52 (0.50)
##   lkl (mean (sd))   3.21 (1.78)
```



UC Berkeley

This site used the RPP Qualtrics file instead of the updated ML5 one. The RPP file had exactly the same wording for the main questions but did not have the new "mechanistic" questions; hence all the missing data that the function complains about.

```
setwd("~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Raw data/
d = read.csv("raw_ucb.csv", header = TRUE)

# merge end-number columns warning about 'NAs introduced by
# coercion', but is correct
d$end.num = coalesce(as.numeric(as.character(d$Q3)), as.numeric(as.character(d$Q8)))
```

```
## Warning in coalesce(as.numeric(as.character(d$Q3)),
## as.numeric(as.character(d$Q8))): NAs introduced by coercion
```

```
## Warning in check_length(val, x, name): NAs introduced by coercion
```

```
# merge effort-split columns warning about 'NAs introduced by
# coercion', but is correct
d$eff.split = coalesce(as.numeric(as.character(d$Q4_1)), as.numeric(as.character(d$Q9_1)))
```

```
## Warning in coalesce(as.numeric(as.character(d$Q4_1)),
## as.numeric(as.character(d$Q9_1))): NAs introduced by coercion
```

```
## Warning in coalesce(as.numeric(as.character(d$Q4_1)),
## as.numeric(as.character(d$Q9_1))): NAs introduced by coercion
```

```
# placeholders for vars not collected
d$badness = NA
d$importance = NA
d$count.eff = NA
d$count.hard = NA

write.csv(d, "manualprep_ucb.csv")
```

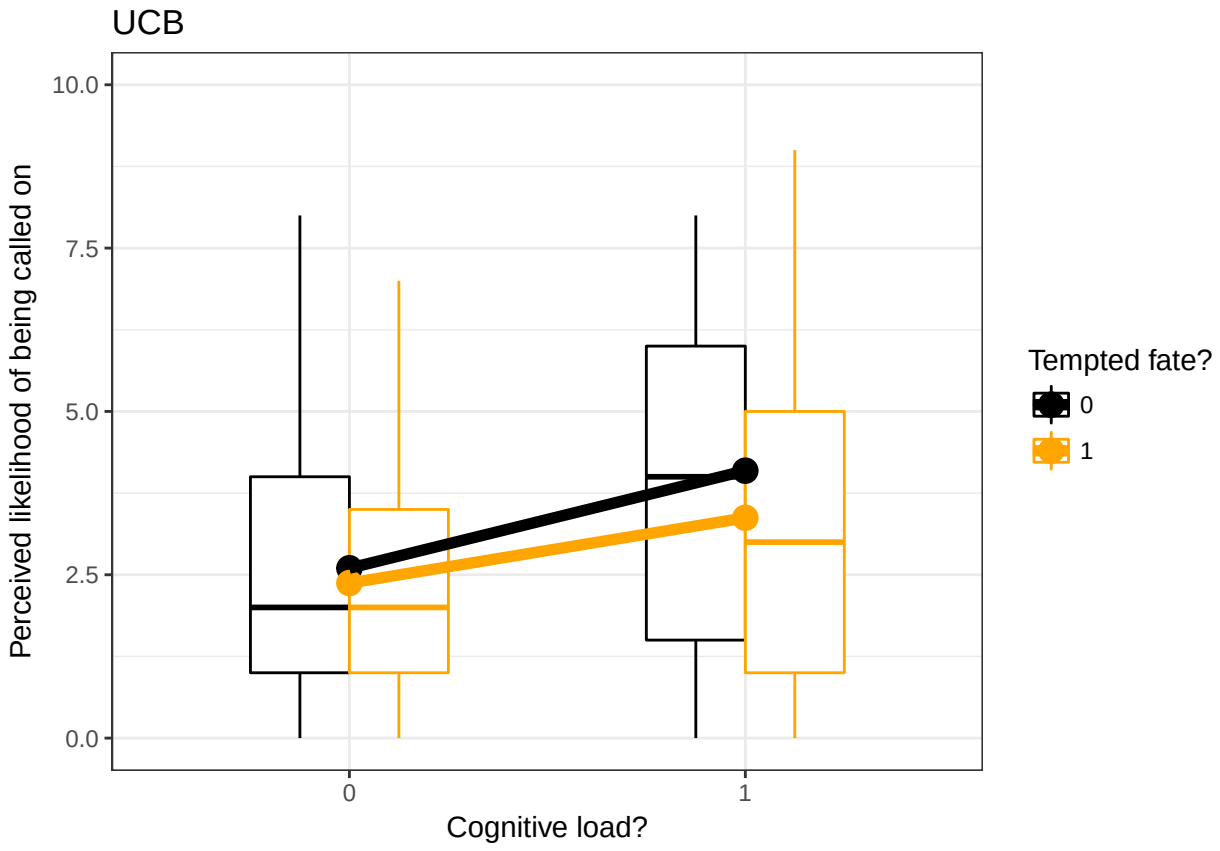
Automatic prep:

```
start.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Ra
end.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Prep

prep_site_data(start.path = start.path, end.path = end.path,
    lkl.names = c("Q2_1", "Q7_1", "Q11_1", "Q14_1"), had.read.name = "Had.read",
    load.name = "Cognitive.load", end.num.name = "end.num", eff.split.name = "eff.split",
    count.eff.name = "count.eff", count.hard.name = "count.hard",
    badness.name = "badness", importance.name = "badness", .site.name = "UCB",
    .group = "b.similar", .n.extra.header.rows = 1)
```

```
##
##
## Extra header rows to delete (first 3 cols):
##   X.1          a mTurkCode
## 1   1 ResponseID mTurkCode
##
##
## First row of real data:
##   X.1                 a mTurkCode           V8           V9 V10 X
## 2   2 R_1cSbvNAXaafqkde   4236033 3/1/17 15:43 3/1/17 16:13   1
##   Cognitive.load Had.read Q1 Q2_1  Q3 Q4_1 Q6 Q7_1 Q8 Q9_1 Q11_1 Q14_1 Q13
## 2              1        1  1    1 549    4
##   end.num eff.split badness importance count.eff count.hard
## 2     549         4      NA         NA        NA         NA
##
## Rows in raw data = 224
##
## Head of skinny dataset before exclusions:
##   id .site.name    .group had.read load lkl eff.split count.eff count.hard
## 1  1        UCB b.similar        1    1   1         4        NA         NA
## 2  2        UCB b.similar        0    1   7         1        NA         NA
## 3  3        UCB b.similar        0    1   3         5        NA         NA
## 4  4        UCB b.similar        1    0   2        NA        NA         NA
## 5  5        UCB b.similar        1    0   6        NA        NA         NA
## 6  6        UCB b.similar        1    0   1        NA        NA         NA
##   badness importance end.num
## 1      NA         NA     549
## 2      NA         NA       4
## 3      NA         NA     551
```

```
## 4      NA         NA      NA
## 5      NA         NA      NA
## 6      NA         NA      NA
##
##
## Subjects with missing had.read, load, or lkl: 69
##
## Subjects with load==1 but missing eff.split, count.eff, or count.hard: 1 2 3 8 9 10 11 12 13 15 18 20
##
## Bad subjects (failed to follow instructions): 12 27 39 41 45 50 53 58 91 95 105 112 119 120 124 125
##
## Final n = 200
##
## MARGINAL MEANS AND SDs FOR ANALYSIS AUDIT
##                      Overall
##   n                   200
##   load (mean (sd))  0.43 (0.50)
##   tempt (mean (sd)) 0.51 (0.50)
##   lkl (mean (sd))   3.02 (2.27)
```

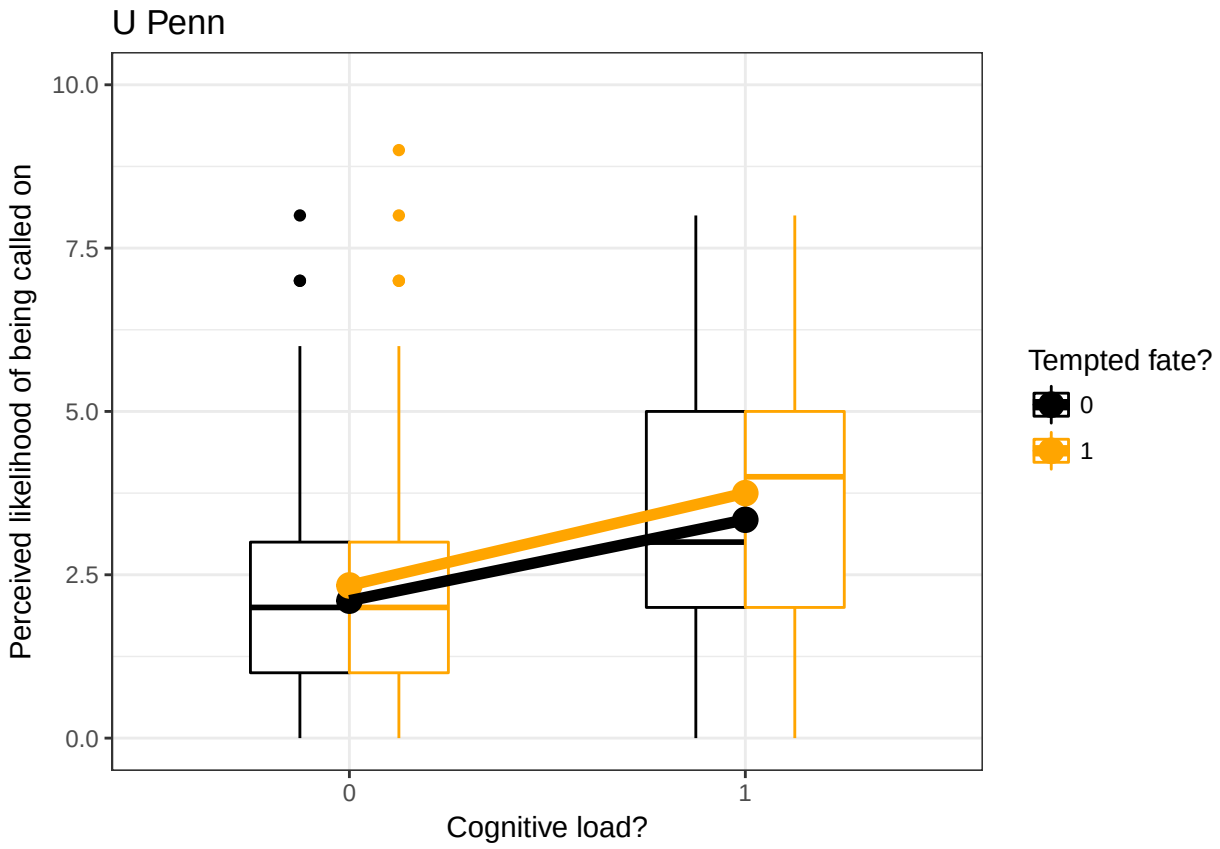



```
start.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Ra
end.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Prep
```

```r
prep_site_data(start.path = start.path, end.path = end.path,
    lkl.names = c("L1.R1.scenario_1", "L1.R0.scenario_1", "L0.R1.text_1",
        "L0.R0.text_1"), had.read.name = "had.read", load.name = "load",
    end.num.name = "Q28", eff.split.name = "Q29_1", count.eff.name = "Q22_1",
    count.hard.name = "Q21_1", badness.name = "Q14_1", importance.name = "Q26_1",
    .site.name = "U Penn", .group = "b.similar", .n.extra.header.rows = 2)
```

```
##
##
## Extra header rows to delete (first 3 cols):
##                   StartDate                 EndDate                  Status
## 1                Start Date                End Date           Response Type
## 2 {"ImportId":"startDate"} {"ImportId":"endDate"} {"ImportId":"status"}
##
##
## First row of real data:
##         StartDate          EndDate     Status     IPAddress Progress
## 3 3/20/2017 10:05 3/20/2017 10:06 IP Address 128.91.96.127      100
##   Duration..in.seconds. Finished    RecordedDate        ResponseId
## 3                    40     TRUE 3/20/2017 10:06 R_1KdMRXffdcx7CRf
##   RecipientLastName RecipientFirstName RecipientEmail ExternalReference
## 3
##   LocationLatitude LocationLongitude DistributionChannel   Q18       Q22
## 3      39.95970154      -75.19680023           anonymous 38717 I Approve
##   L1.R1.scenario_1 L1.R0.scenario_1 Q28 Q29_1 Q21_1 Q22_1 L0.R1.text_1
## 3
##   L0.R0.text_1 Q26_1 Q14_1 load had.read
## 3            3    10    10    0        0
##
## Rows in raw data = 359
##
## Head of skinny dataset before exclusions:
##   id .site.name    .group had.read load lkl eff.split count.eff count.hard
## 1  1     U Penn b.similar        0    0   3        NA        NA         NA
## 2  2     U Penn b.similar        0    0   1        NA        NA         NA
## 3  3     U Penn b.similar        0    0   3        NA        NA         NA
## 4  4     U Penn b.similar        1    0   5        NA        NA         NA
## 5  5     U Penn b.similar        1    0   0        NA        NA         NA
## 6  6     U Penn b.similar        1    0   0        NA        NA         NA
##   badness importance end.num
## 1      10         10      NA
## 2      10         10      NA
## 3      10         10      NA
## 4       2          6      NA
## 5       6          3      NA
## 6       8          8      NA
##
##
## Subjects with missing had.read, load, or lkl:
##
## Subjects with load==1 but missing eff.split, count.eff, or count.hard: 85
##
## Bad subjects (failed to follow instructions): 33 45 46 48 66 100 118 120 125 135 146 169 171 193 211
```

```
##
## Final n = 335
##
## MARGINAL MEANS AND SDs FOR ANALYSIS AUDIT
##                       Overall
##   n                    335
##   load (mean (sd))  0.47 (0.50)
##   tempt (mean (sd)) 0.50 (0.50)
##   lkl (mean (sd))   2.85 (2.03)
```



Mechanical Turk

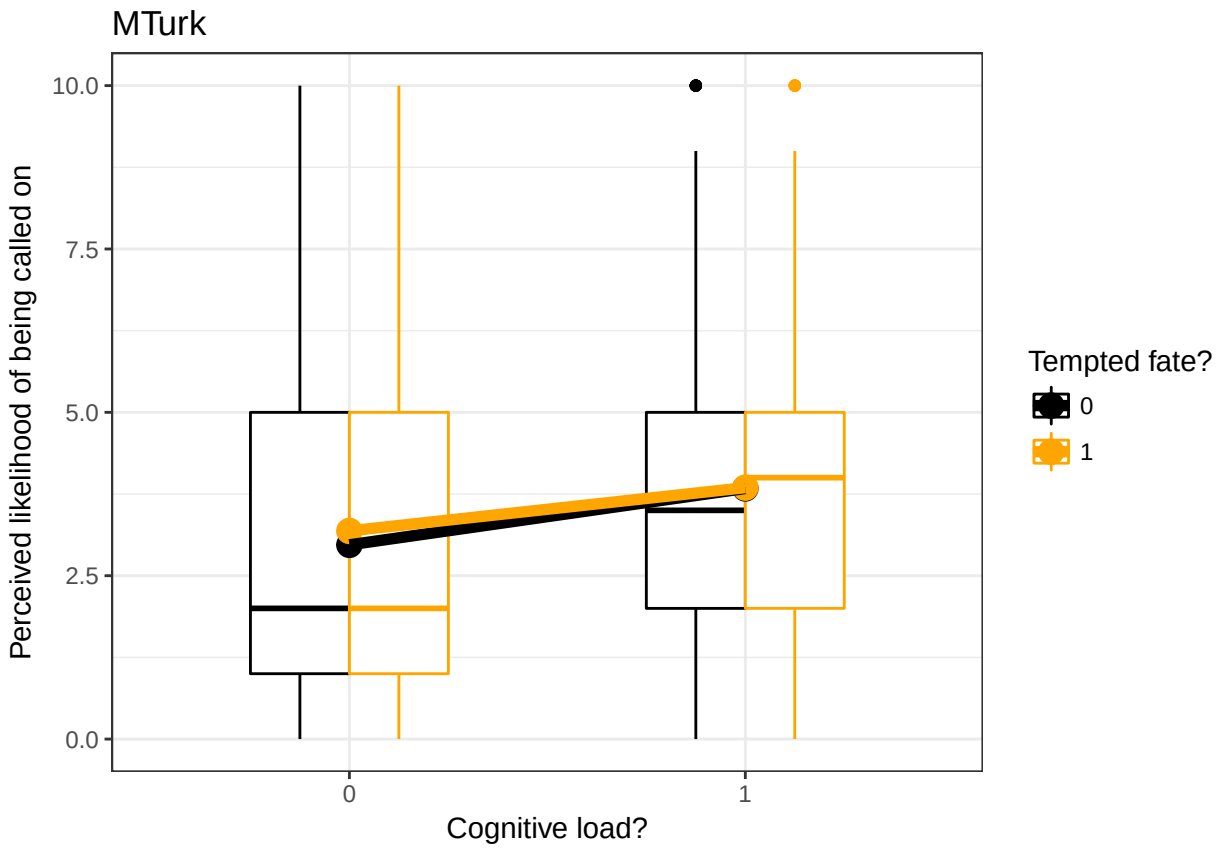
```
start.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Ra
end.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Prep

prep_site_data(start.path = start.path, end.path = end.path,
    lkl.names = c("L1.R1.scenario_1", "L1.R0.scenario_1", "L0.R1.text_1",
        "L0.R0.text_1"), had.read.name = "had.read", load.name = "load",
    end.num.name = "Q28", eff.split.name = "Q29_1", count.eff.name = "Q22_1",
    count.hard.name = "Q21_1", badness.name = "Q14_1", importance.name = "Q26_1",
    .site.name = "MTurk", .group = "a.mturk", .n.extra.header.rows = 2)
```

```
##
##
```

```
## Extra header rows to delete (first 3 cols):
##                  StartDate                EndDate                Status
## 1               Start Date               End Date         Response Type
## 2 {"ImportId":"startDate"} {"ImportId":"endDate"} {"ImportId":"status"}
##
##
## First row of real data:
##      StartDate      EndDate     Status   IPAddress Progress
## 3 6/3/17 15:19 6/3/17 15:20 IP Address 73.0.20.244      100
##   Duration..in.seconds. Finished RecordedDate        ResponseId
## 3                   103     TRUE 6/3/17 15:20 R_3JaOEg0ud90Edcl
##   RecipientLastName RecipientFirstName RecipientEmail ExternalReference
## 3
##   LocationLatitude LocationLongitude DistributionChannel L1.R1.scenario_1
## 3      26.11300659      -80.32559967           anonymous
##   L1.R0.scenario_1 Q28 Q29_1 Q21_1 Q22_1 L0.R1.text_1 L0.R0.text_1 Q26_1
## 3                5 497     2    10    10                                7
##   Q14_1    Q16 Q18                      Q20       Q22 mTurkCode load
## 3     3 Female  25 Graduated 4-year college Caucasian    423773    1
##   had.read
## 3        0
##
## Rows in raw data = 3444
##
## Head of skinny dataset before exclusions:
##   id .site.name  .group had.read load lkl eff.split count.eff count.hard
## 1  1      MTurk a.mturk        0    1   5         2        10         10
## 2  2      MTurk a.mturk        0    0   1        NA        NA         NA
## 3  3      MTurk a.mturk        1    0   1        NA        NA         NA
## 4  4      MTurk a.mturk        1    1   2         5         6          4
## 5  5      MTurk a.mturk        1    0   2        NA        NA         NA
## 6  6      MTurk a.mturk        0    0   1        NA        NA         NA
##   badness importance end.num
## 1       3          7     497
## 2       9          7      NA
## 3       9          5      NA
## 4       6          7     552
## 5       7          8      NA
## 6       8          5      NA
##
##
## Subjects with missing had.read, load, or lkl: 23 724 725 727 728 730 731 732 736 739 740 741 746 747
##
## Subjects with load==1 but missing eff.split, count.eff, or count.hard: 23 725 727 728 730 731 732 735
##
## Bad subjects (failed to follow instructions): 71 116 125 130 139 152 159 170 172 212 263 293 309 317
##
## Final n = 2973
##
## MARGINAL MEANS AND SDs FOR ANALYSIS AUDIT
##                     Overall
##   n                 2973
##   load (mean (sd))  0.44 (0.50)
##   tempt (mean (sd)) 0.50 (0.50)
```

```
##   lkl (mean (sd))   3.42 (2.39)
```



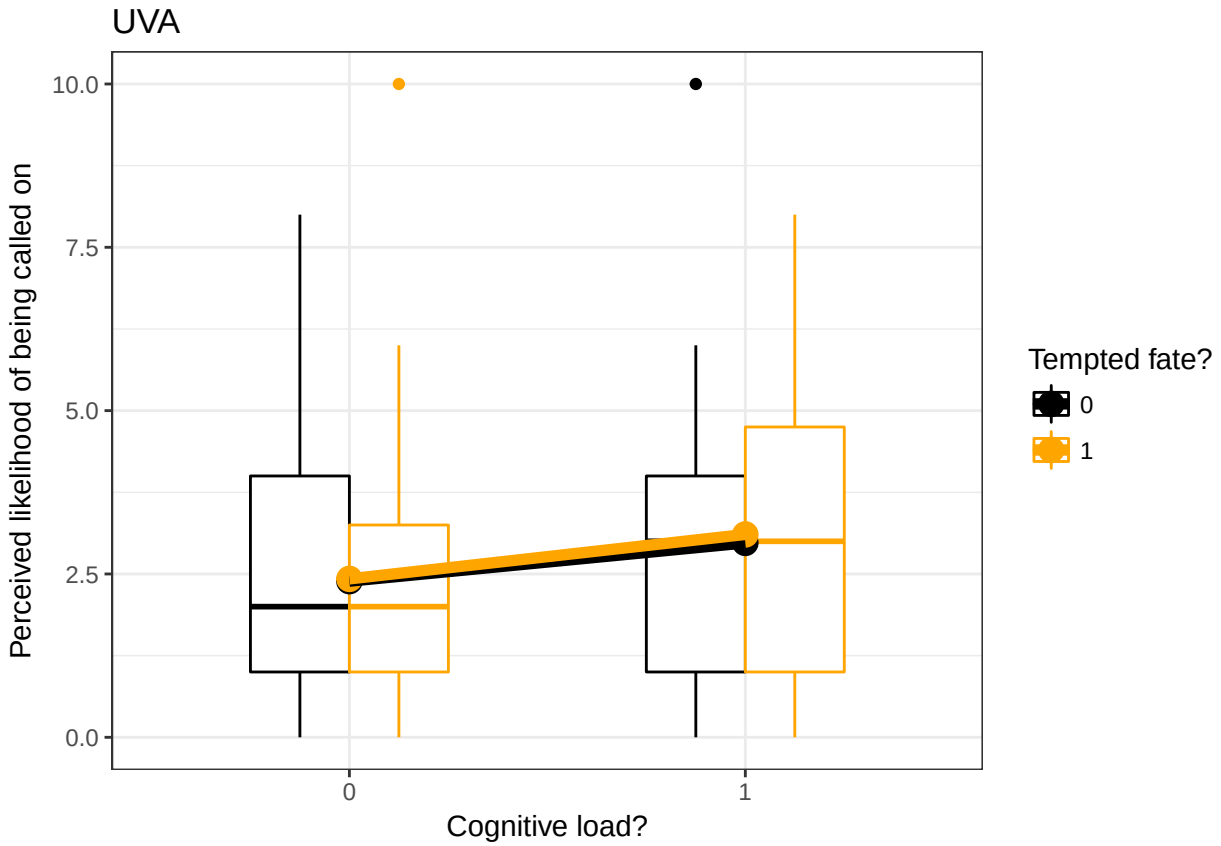
University of Virginia (UVA)

```
start.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Ra
end.path = "~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Prep

d = prep_site_data(start.path = start.path, end.path = end.path,
    lkl.names = c("likelihood_1", "likelihood_1.1", "likelihood_1.2",
        "likelihood_1.3"), had.read.name = "had.read", load.name = "load",
    end.num.name = "end.num", eff.split.name = "effort.split_1",
    count.eff.name = "effort.count_1", count.hard.name = "difficulty_1",
    badness.name = "negativity_1", importance.name = "academic.pressure_1",
    .site.name = "UVA", .group = "b.similar", .n.extra.header.rows = 2)
```

```
##
##
## Extra header rows to delete (first 3 cols):
##                   StartDate                 EndDate                  Status
## 1               Start Date                End Date           Response Type
## 2 {"ImportId":"startDate"} {"ImportId":"endDate"} {"ImportId":"status"}
##
##
## First row of real data:
```

```
##             StartDate             EndDate     Status     IPAddress
## 3 2017-03-21 11:22:11 2017-03-21 11:22:50 IP Address 137.54.25.210
##   Progress Duration..in.seconds. Finished        RecordedDate
## 3      100                    39     True 2017-03-21 11:22:51
##          ResponseId RecipientLastName RecipientFirstName RecipientEmail
## 3 R_3dPVVIn6CnrOyQF
##   ExternalReference LocationLatitude LocationLongitude DistributionChannel
## 3                   38.032897949219  -78.513702392578           anonymous
##   likelihood_1 likelihood_1.1 end.num effort.split_1 difficulty_1
## 3
##   effort.count_1 likelihood_1.2 likelihood_1.3 academic.pressure_1
## 3                             3                                  6
##   negativity_1 mTurkCode load had.read
## 3            7    423601    0        1
##
## Rows in raw data = 165
##
## Head of skinny dataset before exclusions:
##   id .site.name    .group had.read load lkl eff.split count.eff count.hard
## 1  1        UVA b.similar        1    0   3        NA        NA         NA
## 2  2        UVA b.similar        1    1   4         2        10          9
## 3  3        UVA b.similar        0    1   4         4         5          6
## 4  4        UVA b.similar        1    0   4        NA        NA         NA
## 5  5        UVA b.similar        0    0   4        NA        NA         NA
## 6  6        UVA b.similar        1    1   5         5         8          9
##   badness importance end.num
## 1       7          6      NA
## 2       6          8     549
## 3       7          6     543
## 4       8          8      NA
## 5       2          8      NA
## 6       6          7     507
##
##
## Subjects with missing had.read, load, or lkl:
##
## Subjects with load==1 but missing eff.split, count.eff, or count.hard:
##
## Bad subjects (failed to follow instructions): 30 118 119 124 153
##
## Final n = 160
##
## MARGINAL MEANS AND SDs FOR ANALYSIS AUDIT
##                     Overall
##   n                  160
##   load (mean (sd))  0.49 (0.50)
##   tempt (mean (sd)) 0.49 (0.50)
##   lkl (mean (sd))   2.72 (2.11)
```



Aggregated Data Preparation

Stitch datasets:

```
setwd("~/Dropbox/Personal computer/Independent studies/Many Labs 5 (ML5)/Linked to OSF/2. Data/Prepped d

# if there is already prepped data from a previous run,
# remove it to avoid attempting to rbind it with site data
if ("prepped_data.csv" %in% list.files()) file.remove("prepped_data.csv")
```

```
## [1] TRUE
```

```
# rbind all the datasets into one
b = do.call("rbind", lapply(list.files(), function(x) read.csv(x,
    header = TRUE)))

# rename a few cols
names(b)[names(b) == ".site.name"] = "site"
names(b)[names(b) == ".group"] = "group"

# add a few cols
b$is.mturk = 0
b$is.mturk[b$group == "a.mturk"] = 1
```

```r
# write data
write.csv(b, "prepped_data.csv")
```