SGC_3A: The Insight Hypothesis

Primary Analysis for SGC3A

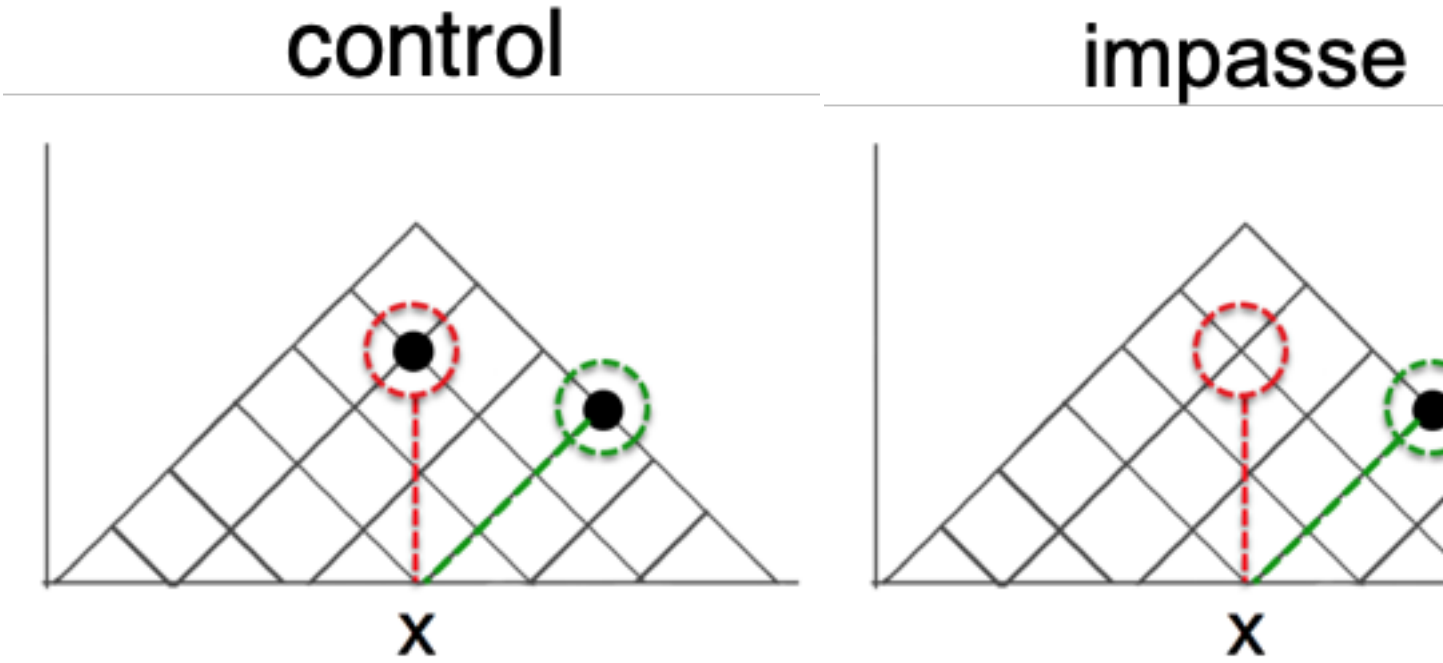
Amy Rae Fox

Contents

INTRODUCTION

In Study 3A we explore a hypothesis that emerged from analysis of Study 2, namely that presenting a learning with a situation that induces a state of impasse will increase the probability they have a moment of insight.

In the context of Study 2, an impasse state was (unintentionally) induced when the combination of question + data set yielded no available answer in the incorrect (cartesian) interpretation of the graph. In Study 3A, we test this hypothesis by comparing performance between a (treatment) group receiving impasse-inducing questions followed by normal questions, and a non-impasse control.



Hypotheses

H1. Learners posed with impasse-inducing questions will be *more likely* to correct interpret the graph.
H0. Learners posed with impasse-inducing questions will be *no more likely* to correctly interpret the graph.

```
#FOR PUBLIC WEB VERSION
# ---
# **To try the study yourself: **
# visit TODO INSERT LINK
# *Enter "github" as your session code, and number of the condition you wish to test*
# session code= GITHUB
# condition code for CONTROL = 111
# condition code for IMPASSE = 121
# <br> <br>
```

METHODS

Design

We employed a mixed design with 1 between-subjects factor with 2 levels (Scaffold: control, impasse) and 15 items (within-subjects factor).

Independent Variables:
- B-S (Scaffold: control,impasse)
- W-S (Item x 15)

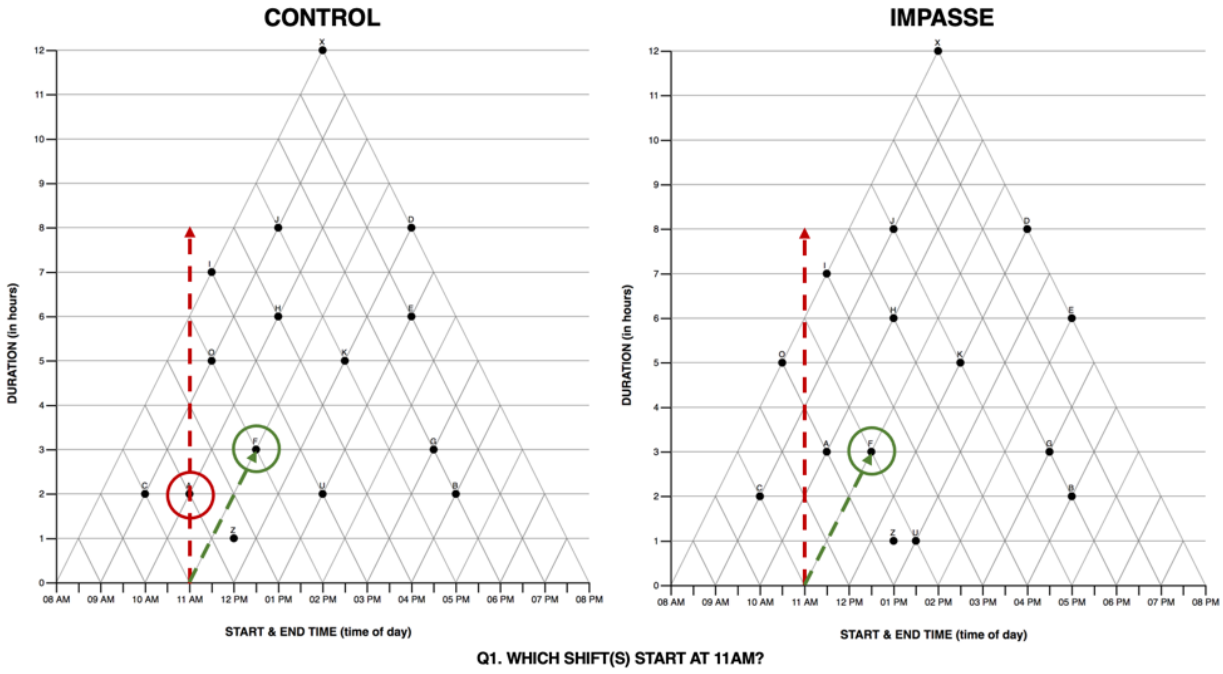
Dependent Variables:
- Response Accuracy : Is the response triangular-correct? 0 (false), 1 (true)
- Response Latency : Time from stimulus onset to clicking 'Submit' button: time in (s)
- (derived) Interpretation : With which interpretation of the graph is the subject's response on an individual question consistent?

TODO: ADD context on scoring algorithm when done

Materials

Stimuli consisted of a series of 15 graph comprehension questions, each testing a different combination of time interval relations, to be read from a Triangular-Model graph. In both conditions, the questions were identical. Participants in the IMPASSE condition saw a graph with a slightly different dataset, such that the some of the questions posed an IMPASSE structure, where no datapoint intersects the orthogonal projection from the x-axis required if reading the graph in accordance with the cartesian coordinate system. The complete list of questions can be found here. Examples of graphs+datasets stimuli for each condition are depicted below.

The green line indicates the ideal-scanpath to the correct (triangular) answer to the first question, and the red line indicates the (incorrect) orthogonal interpretation. In the IMPASSE figure (at right), there are no data points that intersect the red line.



TODO FIX PICTURE SIZE ON PDF

Procedure

Participants completed the study via a web-browser. Upon starting, they submitted informed consent, before reading task instructions. Participants were introduced to a scenario in which they were to play the role of a project manager, scheduling shifts for a group of employees. The schedule of the employees was presented in a TriangularModel (TM) graph, and they would be answering question about the schedule. Then participants completed a test block of 15 items. In the IMPASSE condition, the first five questions included an IMPASSE problem state. For participants in the CONTROL condition, the dataset was structure such that there was always an available 'orthogonal answer' for the first 5 questions. In both conditions, the remaining 10 questions were not structured as impasse. Following the test block, participants answered a free-response question about their strategy for reading the graph, followed by a demographic questionnaire and debrief.

Sample

Data was collected by convenience sample of a university subject pool. Initial data (Fall 2017, Spring 2018) were collected in-person, with large groups of students simultaneously completing the study (independently) in a computer lab. In Fall 2021 and Winter 2022 we collected additional data to replicate results in a remote format (students completing the study asynchronously on their own computers).

DESCRIPTIVES

TODO HARMONIZATION FILE

```
#IMPORT PARTICIPANT DATA
fall17_participants <- "data/fall17_sgc3a_participants.csv"
spring18_partcipants <- "data/spring18_sgc3a_participants.csv"
fall21_participants <- "data/fall21_sgc3a_participants.csv"
winter22_participants <- "data/winter22_sgc3a_participants.csv"

df_fall17 <- read.csv(fall17_participants) %>% mutate(mode = "lab-synch", term = "fall17")
df_spring18 <- read.csv(spring18_partcipants) %>% mutate(mode = "lab-synch", term = "spring18")
df_fall21 <- read.csv(fall21_participants) %>% mutate(mode = "online-asynch", term = "fall21")
df_winter22 <- read.csv(winter22_participants) %>% mutate(mode = "online-asynch", term = "winter22")


#TODO HARMONIZATION FILE THIS IS JUST A TEMPORARY WORKAROUND
#select only columns present in other files

df_subjects <- rbind(df_fall17, df_spring18, df_fall21) %>%
  mutate(
    totaltime_m = totalTime / 1000 / 60,
    absolute_score = triangular_score) %>%
  dplyr::select(subject, condition, session, term, mode, sex, age, totaltime_m, absolute_score)
```

```
df_winter22 <- df_winter22 %>%
  mutate(score = absolute_score, sex = gender) %>%
  dplyr::select( subject, condition, session, term, mode, sex, age, totaltime_m, absolute_score)

df_subjects <- rbind(df_subjects, df_winter22) %>%
  mutate(
    subject = as.factor(subject),
    condition = as.factor(condition),
    session = as.factor(session),
    term = as.factor(term),
    mode = as.factor(mode),
  )

df_lab <- df_subjects %>% filter(mode == "lab-synch")
df_online <- df_subjects %>% filter(mode == "online-asynch")

#Remove extraneous dfs
rm(df_fall17, df_fall21, df_spring18, df_winter22)
```

Participants

```
mode.stats <-  df_subjects %>% group_by(mode,condition) %>% summarize(n=n())

addmargins(table(df_subjects$mode, df_subjects$condition))
```

```
##
##                 111 121 Sum
##   lab-synch      62  64 126
##   online-asynch  96 109 205
##   Sum           158 173 331
```

Data were collected from 126 participants in person, and 205 asynchronously online.

```
#Describe participants

subject.stats <- favstats(age~mode, data = df_subjects)
subject.stats$female <- df_subjects %>% group_by(mode) %>% summarise( female = sum(sex == "Female")/n()
```

For **in-person** collection, 126 participants (60 % female) undergraduate STEM majors at a public American University participated *in person* in exchange for course credit (age: 18 - 33 years). Participants were randomly assigned to one of two experimental groups, with 62 in the control condition, and 64 in the experimental IMPASSE condition.

For **remote, online** replication, 205 participants (70 % female) undergraduate STEM majors at a public American University participated *asynchronously, online* in exchange for course credit (age: 18 - 31 years). Participants were randomly assigned to one of two experimental groups, with 96 in the control condition, and 109 in the experimental IMPASSE condition.

Response Accuracy

Response accuracy refers to how many questions the subject answers with a (strictly correct) triangular interpretation.

```
#DESCRIBE distribution of triangular-correct scores
score.stats <- favstats(absolute_score ~ mode, data = df_subjects)
score.stats
```

```
##            mode min Q1 median Q3 max     mean       sd   n missing
## 1     lab-synch   1  2      3 11  15 5.809524 4.893611 126       0
## 2 online-asynch   0  2      2  8  15 4.668293 4.738993 205       0
```

For *in person* collection, accuracy scores (n = 126) range from 1 to 15 with a mean score of (M = 5.81, SD = 4.89).

For *online replication*, (online) accuracy scores (n = 205) range from 0 to 15 with a mean score of (M = 4.67, SD = 4.74).

Accuracy by Condition

```
#DESCRIBE distribution of triangular-correct scores
lab.scores <- df_lab %>% select(absolute_score, condition) %>% group_by(condition) %>% summarise(summary
print("In Person")
```

```
## [1] "In Person"
```

```
lab.scores
```

```
## # A tibble: 2 x 2
##   condition summary$min   $Q1 $median   $Q3  $max $mean   $sd    $n $missing
##   <fct>           <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <int>    <int>
## 1 111                 1     2     2    3.75    15  4.5   4.64    62        0
## 2 121                 1     2     6.5 12       15  7.08  4.83    64        0
```

```
online.scores <- df_online %>% select(absolute_score, condition) %>% group_by(condition) %>% summarise(s
print("Online")
```

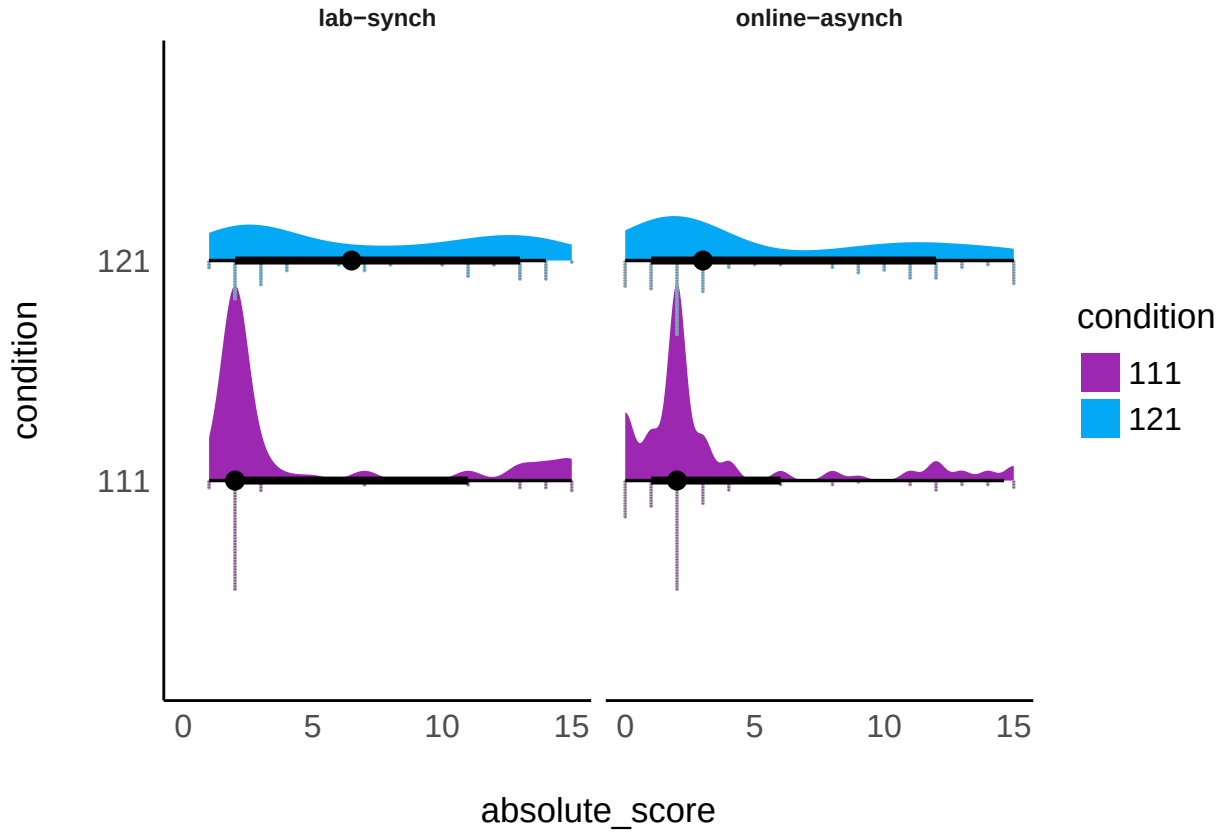
```
## [1] "Online"
```

```
online.scores
```

```
## # A tibble: 2 x 2
##   condition summary$min   $Q1 $median   $Q3  $max $mean   $sd    $n $missing
##   <fct>           <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <int>    <int>
## 1 111                 0  1.75       2     3    15  3.55  4.12    96        0
## 2 121                 0  2          3    11    15  5.65  5.04   109        0
```

```
# #VISUALIZE distribution of response accuracy
# plab <- gf_histogram(~ absolute_score, data = df_lab) +
#   # gf_vline(xintercept = score.stats["lab",]$mean, color = "blue") +
#   labs(title="In-Person")
#
# ponline <- gf_histogram(~ absolute_score, data = df_online) +
#   # gf_vline(xintercept = score.stats["online",]$mean, color = "blue") +
#   labs(title="Online")
#
# plot <-ggarrange(plab, ponline, common.legend = TRUE, nrow = 1, ncol =2)
# annotate_figure(plot, top = text_grob("Score Accuracy by Study",
#    color = "black", face = "bold", size = 14))
```

```
ggplot(df_subjects, aes(y = condition, x = absolute_score, fill = condition)) +
  stat_slab() +
  stat_dotsinterval(side = "bottom", scale = 0.5, slab_size = NA) +
  facet_grid(~mode) +
  theme_modern() +
  scale_fill_material_d(palette = "ice")
```




```
#EASYSTATS SEE
  # geom_violindot(fill_dots = "black") + #ONLY flips dist not dots :/
```

However, inspection of the quantile-quantile plots reveal that response accuracy does not approximate a normal distribution. After exploring several transformations and comparing

against alternative distributions (log-normal, poisson, exp, nbinom), we conclude that we will need to use robust tests to analyze response accuracy.

```
# plot(fitdist(df_subjects$absolute_score, "norm"))

#EXPLORE ALTERNATIVES
# fit_n  <- fitdist(df_fall$absolute_score, "norm")
# fit_p  <- fitdist(df_fall$absolute_score, "pois")
# fit_b  <- fitdist(df_fall$absolute_score, "nbinom")

# par(mfrow=c(2,2))
# plot.legend <- c("normal", "poisson","nbinomial")
# denscomp(list(fit_n, fit_p, fit_b), legendtext = plot.legend)
# cdfcomp (list(fit_n, fit_p, fit_b), legendtext = plot.legend)
# qqcomp  (list(fit_n, fit_p, fit_b), legendtext = plot.legend)
# ppcomp  (list(fit_n, fit_p, fit_b), legendtext = plot.legend)
```

Response Latency

```
#DESCRIBE distribution of response time
time.stats <- rbind(
  "lab"= favstats(df_lab$tri_min),
  "online"= favstats(df_online$tri_min)
)
time.stats <- time.stats %>% dplyr::select(-missing) #don't need missing column
time.stats
```

```
##        min Q1 median Q3 max mean sd n
## lab     NA NA     NA NA  NA  NaN NA 0
## online  NA NA     NA NA  NA  NaN NA 0
```

For *in person* response latency (for test block) (n = 0) range from NA to NA minutes, with a mean duration of (M = NaN, SD = NA) minutes.

For *online replication* (online) response latency (for test block) (n = 0) range from NA to NA minutes, with a mean duration of (M = NaN, SD = NA).

```
# #VISUALIZE distribution of response time
# plab <- gf_dhistogram(~tri_min, data = df_lab) %>%
#   # gf_vline(xintercept = time.stats["lab",]$mean, color = "black") %>%
#   # gf_fitdistr(color="red")+
#   labs(title="In Person")
#
# ponline <- gf_dhistogram(~tri_min, data = df_subjects) %>%
#   gf_vline(xintercept = time.stats["online",]$mean, color = "black") %>%
#   gf_fitdistr(color="red")+
#   labs(title="Online")
#
# plot <-ggarrange(plab, ponline, common.legend = TRUE, nrow = 1, ncol =2)
#
# annotate_figure(plot, top = text_grob("Total Time by Study",
#    color = "black", face = "bold", size = 14))
```

The data may need to be log-transformed. But we will address this when modelling with the variable.

TODO ADD ITEM LEVEL

HYPOTHESIS TESTING

Response Accuracy by Condition

The experimental hypothesis (H1) is that structuring the data to pose an impasse (condition 121) will produce significantly better performance than non-impasse (condition 111). The null hypothesis (H0) is that there will be no difference in performance between conditions.

[EXPLORE]

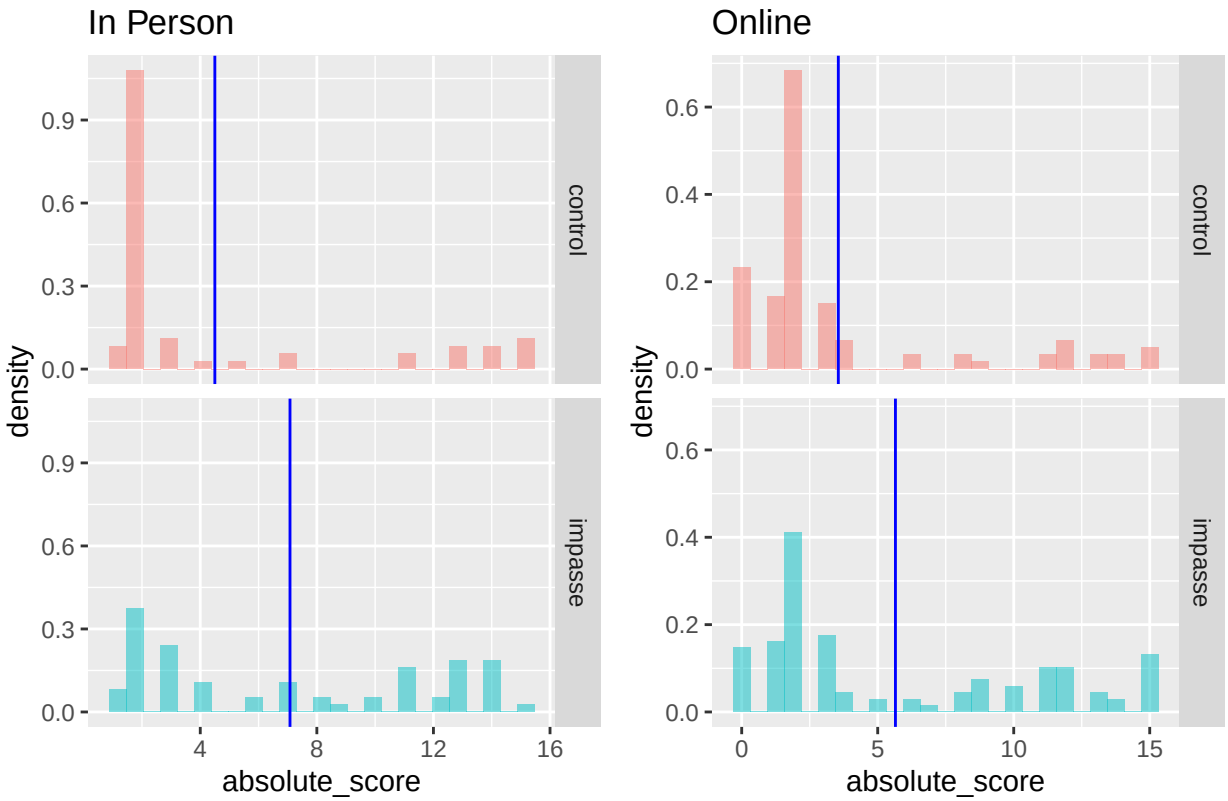
```
#DESCRIBE scores by condition
score.cond.stats <- rbind(
  "lab" = favstats(absolute_score ~ condition, data = df_lab),
  "online" = favstats(absolute_score ~ condition, data = df_online)
)
score.cond.stats
```

```
##          condition min   Q1 median    Q3 max     mean       sd   n missing
## lab.1          111   1 2.00    2.0  3.75  15 4.500000 4.643875  62       0
## lab.2          121   1 2.00    6.5 12.00  15 7.078125 4.828174  64       0
## online.1       111   0 1.75    2.0  3.00  15 3.552083 4.118942  96       0
## online.2       121   0 2.00    3.0 11.00  15 5.651376 5.041267 109       0
```

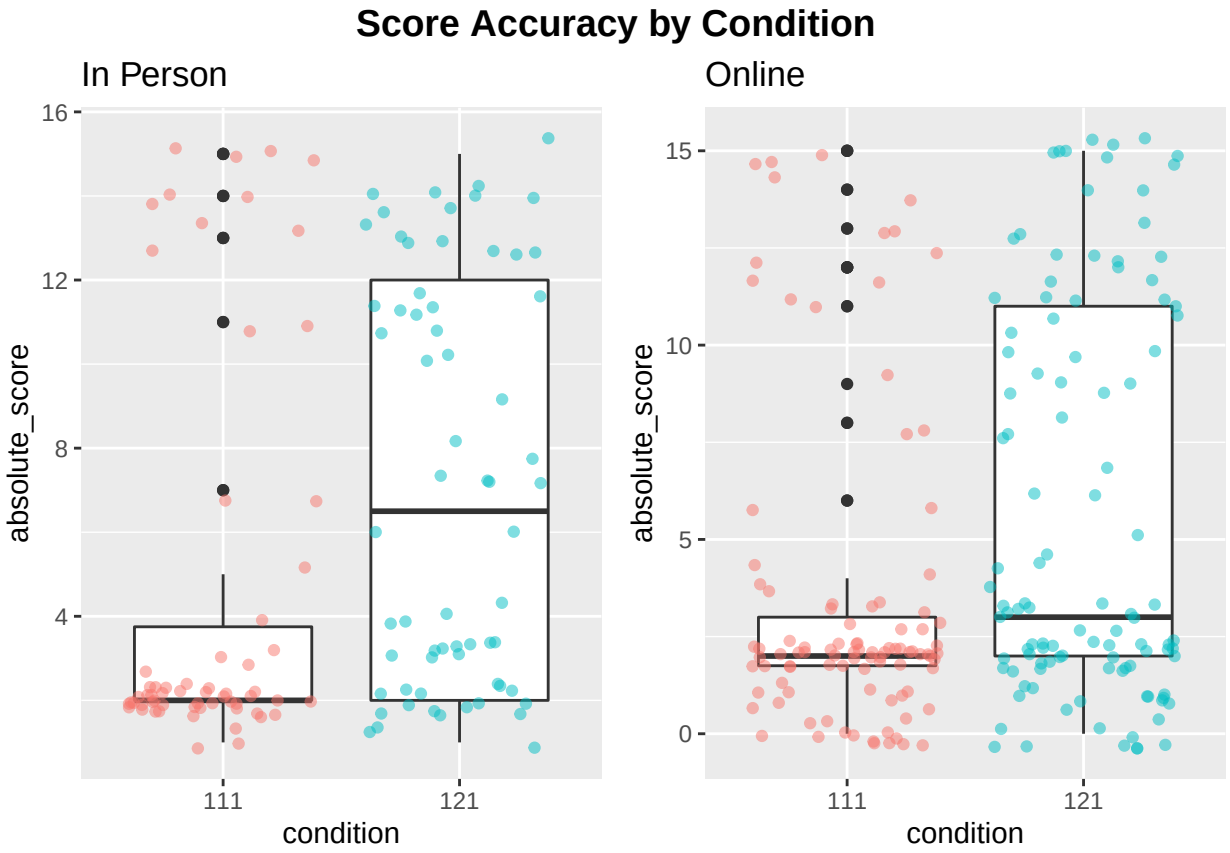
For **in person** study, participants in the impasse group had (on average) higher scores (M = 7.08 SD = 4.83) than those in the non-impasse control group (M = 4.5, SD = 4.64).

For **online replication** study, participants in the impasse group had (on average) higher scores (M = 5.65 SD = 5.04) than those in the non-impasse control group (M = 3.55, SD = 4.12).

```
#VISUALIZE scores by condition
condlables <- c("111"="control", "121"="impasse")
plab <- gf_dhistogram( ~absolute_score, fill= ~condition, data = df_lab) %>%
  gf_facet_grid(condition~., labeller=labeller(condition=condlables)) %>%
  gf_vline(xintercept = ~mean, data = score.cond.stats[c(1:2),], color = "blue")+
  labs(title="In Person")

ponline <- gf_dhistogram( ~absolute_score, fill= ~condition, data = df_online) %>%
  gf_facet_grid(condition~., labeller=labeller(condition=condlables)) %>%
  gf_vline(xintercept = ~mean, data = score.cond.stats[c(3:4),], color = "blue")+
  labs(title="Online")

plot <-ggarrange(plab, ponline, legend = FALSE, nrow = 1, ncol =2)
annotate_figure(plot, top = text_grob("Score Accuracy by Condition",
   color = "black", face = "bold", size = 14))
```




```
#VISUALIZE scores by condition
plab <- gf_boxplot(absolute_score ~ condition, data=df_lab) %>%
  gf_jitter(color=~condition, alpha=0.5) +
  labs (title = "In Person")

ponline <- gf_boxplot(absolute_score ~ condition, data = df_online) %>%
  gf_jitter(color=~condition, alpha=0.5)+
  labs(title ="Online")

plot <-ggarrange(plab, ponline, legend = FALSE, nrow = 1, ncol =2)
annotate_figure(plot, top = text_grob("Score Accuracy by Condition",
   color = "black", face = "bold", size = 14))
```



[MODEL]

Because the response accuracy data are not-normal, we will test the veracity of our hypothesis using Wilcoxon rank-sum test (Wilcoxon, 1945).

```
#Wilcoxon Rank Sum Test
m1 <- wilcox.test(absolute_score ~ condition, data = df_lab, exact=FALSE,
                  alternative="less")
m1
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  absolute_score by condition
## W = 1284.5, p-value = 0.0001922
## alternative hypothesis: true location shift is less than 0
```

```
library(rstatix)
#Calculate Effect size
m1eff <- wilcox_effsize(absolute_score ~ condition, data = df_lab)
m1eff
```

```
## # A tibble: 1 x 7
##   .y.            group1 group2 effsize    n1    n2 magnitude
```

```
## * <chr>          <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 absolute_score 111    121      0.317    62    64 moderate
```

In Person A one-tailed Wilcoxon rank-sum test reveals that the cummulative score (number of triangular-consistent responses)(Mdn = 6.5 points) in the IMPASSE group were significantly higher than than scores in the non-impasse control condition (Mdn = 2), W = 1284.5 , p < 0.001, *r* = 0.3165359, a moderate-sized effect.

```
#Wilcoxon Rank Sum Test
m1 <- wilcox.test(absolute_score ~ condition, data = df_online, exact=FALSE,
                  alternative="less")
m1
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  absolute_score by condition
## W = 3994, p-value = 0.001424
## alternative hypothesis: true location shift is less than 0
```

```
library(rstatix)
#Calculate Effect size
m1eff <- wilcox_effsize(absolute_score ~ condition, data = df_online)
m1eff
```

```
## # A tibble: 1 x 7
##   .y.            group1 group2 effsize    n1    n2 magnitude
## * <chr>          <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 absolute_score 111    121      0.208    96   109 small
```

Online A one-tailed Wilcoxon rank-sum test reveals that the cummulative score (number of triangular-consistent responses)(Mdn = 3 points) in the IMPASSE group were significantly higher than than scores in the non-impasse control condition (Mdn = 2), W = 3994 , p < 0.05, *r* = 0.20848, a small effect.

TODO:: Investigate outliers in online study; may need to have stricter response time + strategy-consistent response criteria

[REPLICATION]

Because the effect of impasse scaffold was smaller in the remote-online sample than in-person sample, we'll test whether scores in in-person sample were significantly smaller than those online using another Wilcoxon rank-sum test.

```
temp <- df_subjects %>% filter(condition==121)

#Wilcoxon Rank Sum Test
m1 <- wilcox.test(absolute_score ~ mode, data = temp, exact=FALSE,
                  alternative="greater")
m1
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  absolute_score by mode
## W = 4226, p-value = 0.009639
## alternative hypothesis: true location shift is greater than 0
```

```
#Calculate Effect size
m1eff <- wilcox_effsize(absolute_score ~ mode, data = temp)
m1eff
```

```
## # A tibble: 1 x 7
##   .y.            group1    group2        effsize    n1    n2 magnitude
## * <chr>          <chr>     <chr>           <dbl> <int> <int> <ord>
## 1 absolute_score lab-synch online-asynch   0.178    64   109 small
```

A one-tailed Wilcoxon rank-sum test reveals that the cummulative score (number of triangular-consistent responses)(Mdn = 6.5 points) in the IN-PERSON (impasse condition) sample *were not* significantly higher than than scores in the REMOTE-ONLINE (impasse condition) sample (Mdn = 3), W = 4226 , p = 0.01 r = 0.1780372.

Inference Our replication study had comparable results to the in-person study, however the effect size was smaller, therefore we should consider increasing the sample size of future studies conducted online so as to ensure we have sufficient statistical power detect a smaller effect.

DILLIGENCE

Assumptions of Wilcoxon Rank-Sum

The Wilcoxon rank-sum test is the non-parametric alternative to a independent samples t-test and requires the following assumptions:
1. The two samples are independent of one another <- MET by random sampling + assignment
2. The two populations have equal variance or spread <-TEST

```
df_subjects %>% group_by(condition,mode) %>% summarize(var=var(absolute_score))
```

```
## # A tibble: 4 x 3
## # Groups:   condition [2]
##   condition mode            var
##   <fct>     <fct>         <dbl>
## 1 111       lab-synch      21.6
## 2 111       online-asynch  17.0
## 3 121       lab-synch      23.3
## 4 121       online-asynch  25.4
```

Variances between conditions in both Lab and Online samples are comparable

REPRODUCABILITY

For data dictionary, see **TODO**