Welcome back (+ midterm review)

18 March 2020

Modern Research Methods

Logistics

- As a reminder, this session will be audio/video recorded for educational use by other students in this course.
- Getting to the classroom (links in Canvas and on website)
- Office hours – same time, but over Zoom
- Must **sign-up** using spreadsheet
  - If you're unable to make those times, let us know and we will do our best to accommodate you

INSTRUCTOR

Dr. Molly Lewis

mollylewis@cmu.edu

ZOOM OFFICE

Office Hours: W 4:30-6:30pm

Signup: **signup sheet**

TA

Jaeah Kim

jaeahk@andrew.cmu.edu

ZOOM OFFICE

Office Hours: M 1:00-3:00pm

Signup: **signup sheet**

COURSE

MW (lecture); F (lab)

10:30-11:20am

Lecture/Lab: **ZOOM CLASSROOM**

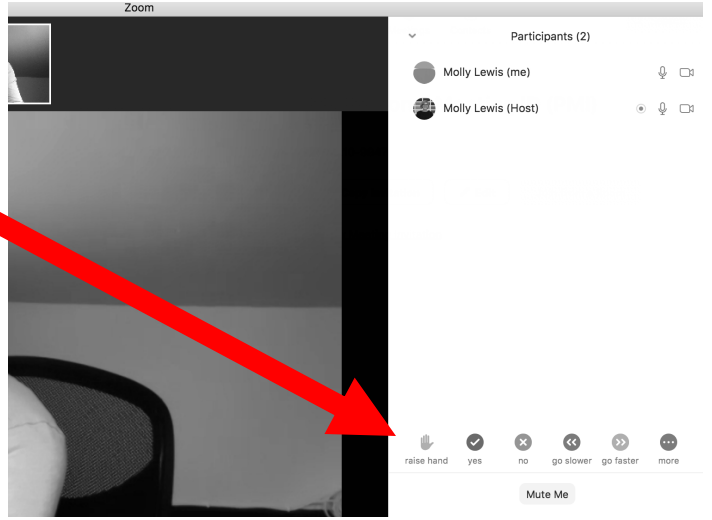
Zoom etiquette

- Mute yourself when you're not actively speaking
  - Unmute to speak
  - Then, mute again after you speak

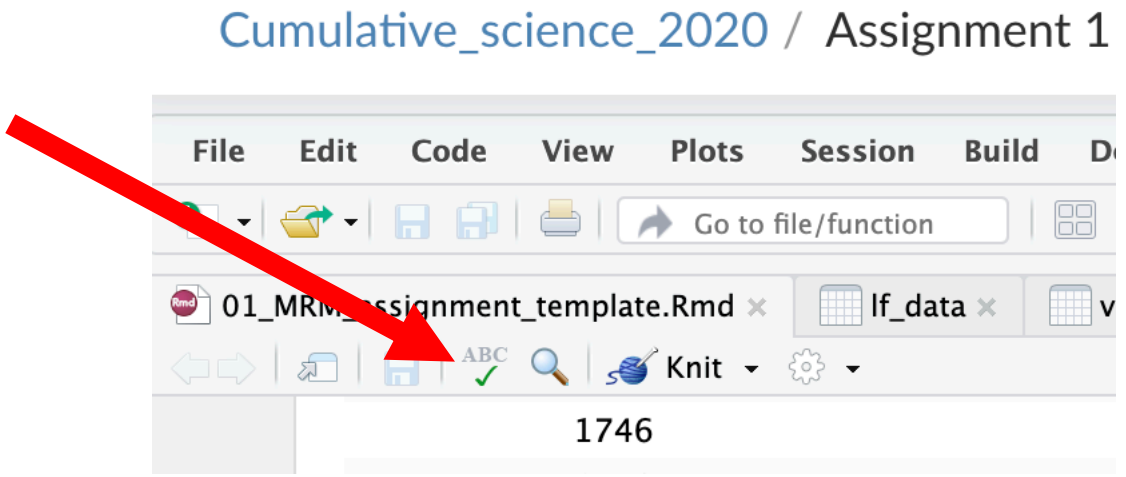
(or hold spacebar to unmute)

- Use headphones if possible
  - Prevents feedback from your speakers

- When you want to speak, raise your hand
  - When I see your hand raised, I'll call on you



Midterm



- Grades are posted
- Spell check!
- Still some formatting issues
  - 5pts on the midterm for appropriately formatted responses
  - If you got penalized for this, there should be a note in canvas
  - You may fix your midterm and turn it back into us to get this points back (by next Wednesday)
- Formatting issues:
  - Printing out whole dataset
  - Not using any headers
  - Writing responses as comments in R code

What does the rest of the semester look like?

- Today, go over midterm
- On Friday, more practice with effect sizes
- Then, we're going to start working on the final project – a meta-analysis
- We'll talk more about this next week, but start thinking about what psychology questions you might be interested in studying

Question 1 What does the term “cumulative science” mean? Your response should make reference to examples we’ve discussed in class and the course readings. [1-2 paragraphs]

- New scientific knowledge is built upon on the findings in previous research and each other’s research
- Involves reproducibility; tools like R markdown and Github facilitate this
- Involves replicability; p-hacking and other QRPs slow cumulative science

Question 2

[a] Define the terms “reliability” and “validity.”
[b] Assess the validity of Zettersten and Lupyan (2020), Exp. 1A.
[c] Describe how you would assess the reliability of Zettersten and Lupyan (2020), Exp. 1A.

[a]

> Reliability is the consistency of a measurement, meaning how much of the time a variable will be reported as the same given multiple measurements of an identical input. There are several subtypes of validity, but in general it refers to how well a measurement actually captures the construct it supposedly measures. For example, how often a person smiles has low validity as a measurement of happiness, because there are many reasons why people smile other than genuine happiness. -Isobel Stephen

[b]

> The experiment set out to test if having a name for something made it easier to categorize. This is somewhat valid for experiment 1A as it was testing easily-named colors vs difficult-to-name colors, but I would say that it doesn’t really extend beyond the scope of color categorization. It does a great job as testing the effects of high-nameability vs. low-nameability on the categorization of color. - Zoe Marshall

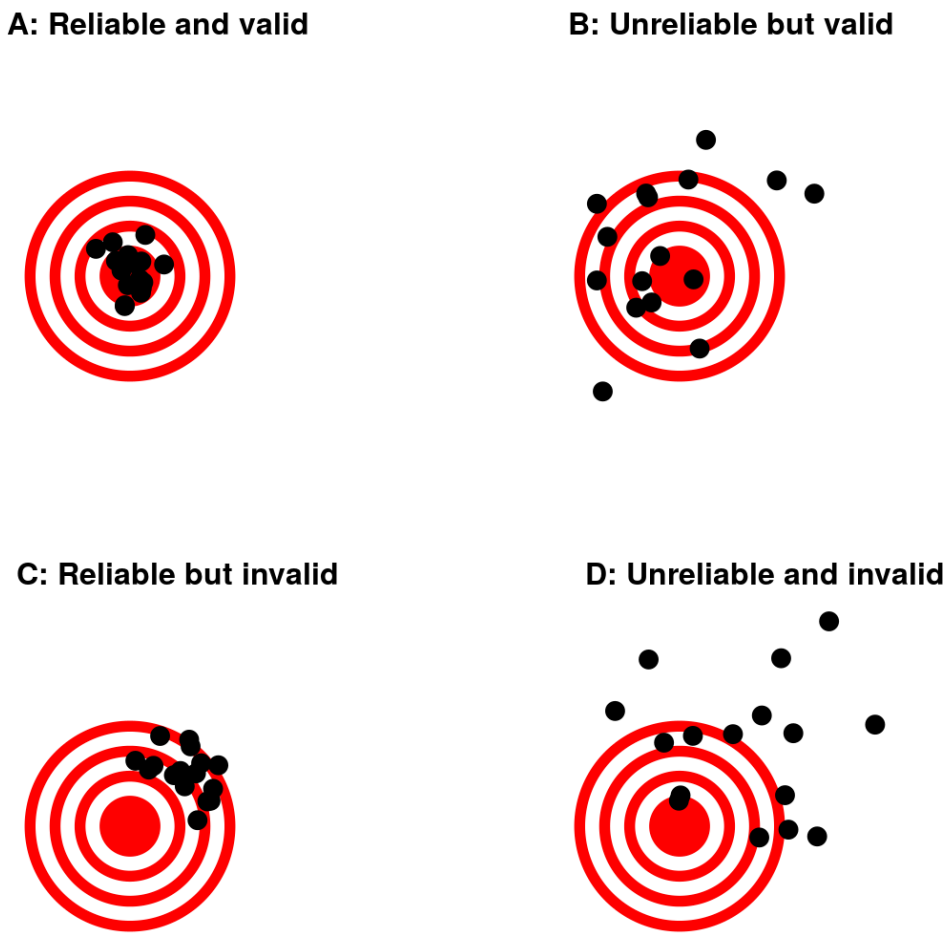
[c] I would assess it by running the experiment again. “Test-retest reliabilty” = running the experiment again with the same subjects and comparing subjects individual performance on trial1 vs. trial2.

What makes a good measurement?

Reliability – Consistency of measurement (Do you say the same thing today as yesterday?)

Validity – Are we measuring the construct we want to measure?

Suppose the thing we're trying to measure is the center of the bullseye…



Question 3

Suppose you were interested in estimating how many words a child knows (their vocabulary size). Describe three different variables that you could measure to estimate the size of a child's vocabulary size (each variable should be of a different type). For each variable, give a one sentence description of the variable, the variable type, *and* one example value of that variable with units.

One qualitative variable that could be used to estimate how many words a child knows would be how sure they are of the meaning of the word with values ranging from "unsure", "moderately sure", and "very sure"; for instance, if asked about how sure they are about the definition of a word, the child could respond by saying that they were "very sure" about their response. - Sarah Chen

Time to list 100 words. (Real Number) 300.21 Seconds.- Funglun Chan

Quantitative variable, Integer: CDI-Score. Calculate the child's score on the Mac-Arthur Bates Communicative Development Inventories. Example: 58, 63…. - Anjie Cao

To estimate the size of a child's vocabulary you could have the variable of the child's age (continuous), level of parent education (categorical), and number of hours books in the house (discrete). The child's age could be measured continuously starting with birth as the meaningful zero point, so a child could be 7.25 years old. The Level of parent education could be coded with ordinal levels such as some high school, high school diploma, undergraduate degree, graduate degree. The number of books in the house is a discrete variable and I think it is useful because other studies have shown that children that are read to more often perform better on language tests later on. An example of a value of this variable could be 33 books. - Nicole Casey

Question 4 Use Google Sheets to look at this dataset.

[a] List 5 things about the data in Table 1 that are not tidy.
[b] What are the observations in this dataset?

If it's helpful, you can learn more about this dataset here

[a]

> Five things that are not tidy in Table 1 are: 1. Year (a general variable) is not its own column; instead, each specific year has a column 2. The international migrant stock at mid-year by sex is a large merged cell acting as a "title" rather than having a column for the numerical variable "international migrant stock at mid-year" and another column for the variable "Sex of international migrant stock" (Male, Female, Both sexes) 3. Geographic regions (such as "Southern Africa") for countries are included in the same column as the country names themselves, when there could be a separate column for geographic region (Benin…..Western Africa, Botswana….Southern Africa, etc.) 4. The values for migrant count in each major area, region, country, or area of destination are all in the same row - labeled by country - rather than each having a distinct row for each observation (One row for international migrant count of females in Nigeria in 1990, one row for international migrant count of males in Nigeria in 1990, one row for international migrant count of both sexes in Nigeria in 1990, one row for international migrant count of females in Nigeria in 1995, one row for international migrant count of males in Nigeria in 1995, etc.) 5. "Summary" data that are larger observations (Migrant counts of people by income group or by UN development groups, rather than by country) are included in the same table when they should be placed into separate tables (Table 1: International migrant stock at mid-year by income group, Table 2: International migrant stock at mid-year by UN development group, Table 3: International migrant stock at mid-year by geographic region, Table 4: International migrant stock at mid-year by country) - Victoria Shiau

[b]

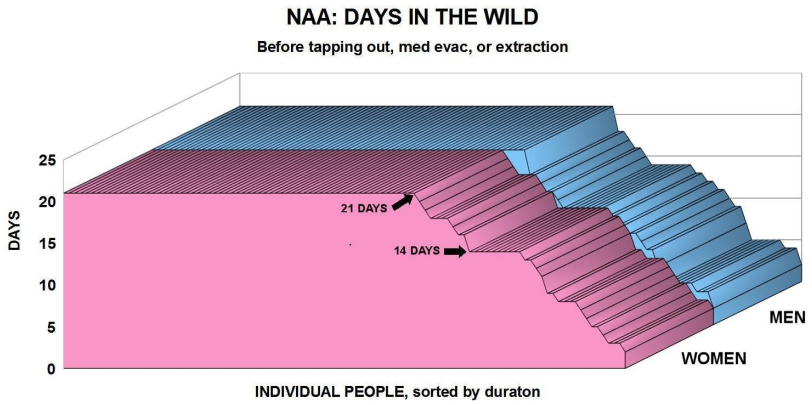
> Observations are the population of international migrant stock of a certain location, sex, year. - Funglun Chan

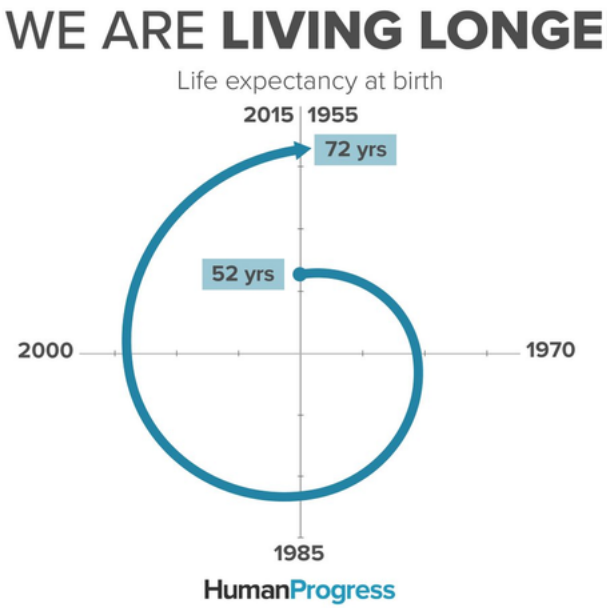
> An observation is the immigrant stock in a given year, for a given gender, in a given country/country group - Yinxuan Fu

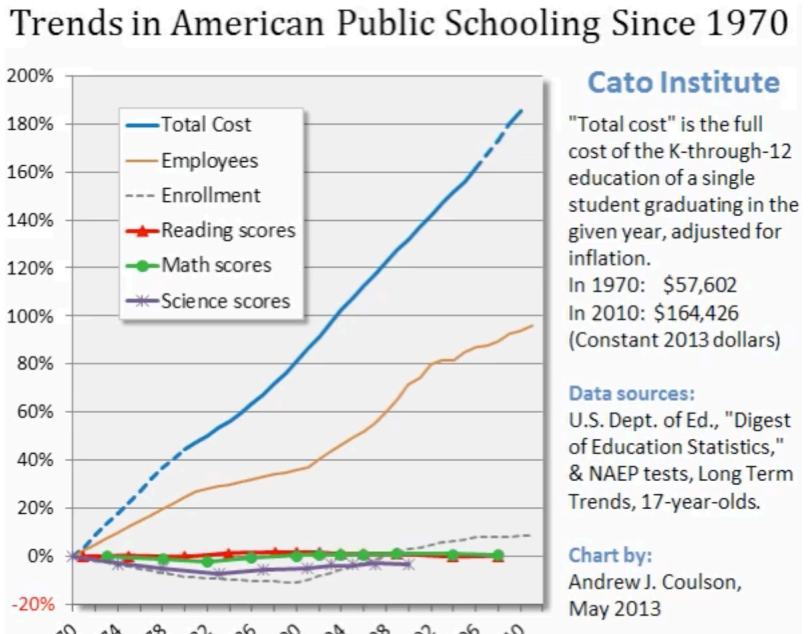
Question 5

Go to the dataisugly subreddit and find one plot that you think is a particularly severe offender of the plotting guidelines we discussed in class.

[a] Provide a link to that plot.
[b] Describe what you take to be the main message of the plot.
[c] Describe 5 things you would change in the plot to make it better at conveying this message.
[d] Imagine you were going to recreate and/or improve upon this plot using `ggplot`. List the name of one geom function you would use to make your plot (e.g. `geom_boxplot`).







Question 6

Rhoda is a researcher in the psychology department at CMU. She has the hypothesis that children will be more likely to show the mutual exclusivity effect in word learning if the child is asked to find the novel object in infant directed speech, compared to adult directed speech. Rhoda runs a study testing this hypothesis with a sample of children from the local nursery school.

[a] Explain what it would mean to reproduce this study.
[b] Explain what it would mean to replicate this study.

a: To reproduce this study would be to get the same results as Rhoda while using the same data set. This would be done by an outside researcher, who would use the same raw data and the same statistical analysis and get the same results. b: To replicate this study would be to repeat the entire experiment, use the same analysis plan, and get the same data pattern as Rhoda. - Themi Bournias

a: To reproduce this study, we would take Rhoda's data and follow her process for data analysis to see if we get the same statistical results as her; b: To replicate this study, we would follow Rhoda's whole experimental process in which we use the same population (children from nursery school), test the same hypothesis, follow the same experimental design and analysis plan, and then see if we get the same result. - Shruti Murali

Question 7

Jasmina conducts a replication of an experiment. She uses her data to calculate (i) a p-value, (ii) confidence intervals, and (iii) an effect size. The paper describing the original experiment also reports a p-value, confidence intervals, and an effect size. Explain what each of these three statistical tools will tell Jasmina about the relationship between the original experiment and her replication.

> p-value: p-value refers to the probability of getting two sample means that are at least as different as the difference we found if the null hypothesis is true. If the p-values in the replication and the original experiment are close to each other and are both small, then it suggests that the replication is successful; Confidence intervals: Confidence intervals refer to a range of plausible values. For example, for a 95% intervals, if we run the experiment over and over again by drawing samples from the same populations and calculate the time we run the experiment, then 95% of the intervals would contain the true mean for the population. In Jasmina's case, we should expect to see overlapping confidence intervals between the original experiment and the replication if the replication is successful. Effect size: Effect size is a standardized way to measure the difference between performance in two conditions and taking into account the dispersion of the two conditions. Larger dispersion will lead to smaller effect size and smaller dispersion will lead to larger effect size. In Jasmina's case, for successful replication, we should also expect to see a similar effect size if the replication has a similar sample size with the original. - Anjie Cao

- P-value: binary statement about whether there's effect [significant/not significant]
  - note: p-value doesn't tell you about the size of the effect since the p-value depends on the confidence interval
- Confidence interval: Range of plausible values for the mean
- Effect size (e.g. Cohen's d): magnitude of the difference between the two means

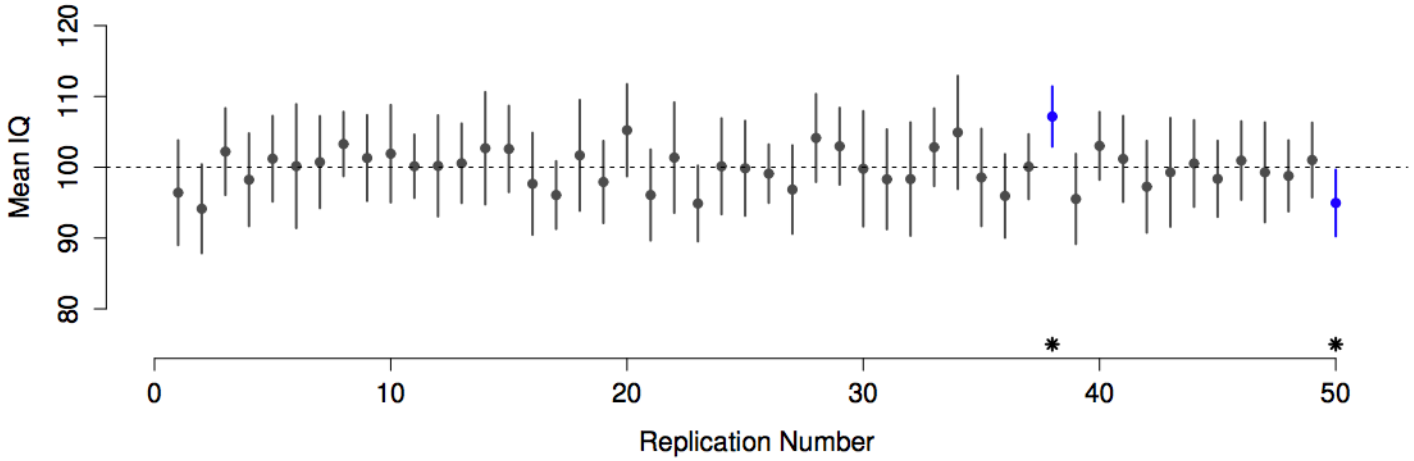
Question 8

The mean IQ of *all the people* on the island Lorbia is 135 (they're very good at taking IQ tests!). You take a sample of 10 people from Lorbia and calculate the mean IQ and a 95% confidence interval around the mean. You repeat this procedure for a total of 200 samples. How many of your 200 confidence intervals will include the value 135?

> .95 * 200 = 190
>
> 190 confidence intervals will contain the true mean. - Jailyn Zabala

Confidence intervals:

- Set of plausible values for an estimate
- If you run an experiment 100 times, the true population value will be included in 95 of the confidence intervals



Question 9

This question uses the data from the Many Babies project from Assignment 5.

[a] Calculate effect sizes for the replications in two of the labs. Specifically, calculate an effect size for the replication experiment conducted by the "babylabnijmegen" lab and an effect size for the replication experiment conducted by the "infantllmadison" lab.
[b] Plot the two effect sizes.
[c] Which lab had the largest effect size? State the point estimate and confidence interval for the effect size for that lab.
[d] Interpret the effect sizes. Explain what it means to have a large effect size in this experiment. Would you guess that these two effect sizes are statistically different from each other? Why or why not?
[e] Imagine we ran another replication of this experiment and got a *negative* effect size. What would this mean?

```
many_babies_data <- read_csv("data/many_babies_data.csv")

many_babies_data_summary <- many_babies_data %>%
  group_by(lab, condition) %>%
  summarize(mean_prop_right = mean(mean_looking_time),
            sd = sd(mean_looking_time),
            n = n())
```

```
babylabnijmegen_effect_size <-
  mes(many_babies_data_summary %>% filter(lab == "babylabnijmegen", condition == "IDS") %>% pull(mean_prop_right), # m.1
    many_babies_data_summary %>% filter(lab == "babylabnijmegen", condition == "ADS") %>% pull(mean_prop_right),# m.2
    many_babies_data_summary %>% filter(lab == "babylabnijmegen", condition == "IDS") %>% pull(sd), #sd.1
    many_babies_data_summary %>% filter(lab == "babylabnijmegen", condition == "ADS") %>% pull(sd), #sd.2
    many_babies_data_summary %>% filter(lab == "babylabnijmegen", condition == "IDS") %>% pull(n), #n.1
    many_babies_data_summary %>% filter(lab == "babylabnijmegen", condition == "ADS") %>% pull(n), #n.2
    verbose = F)  %>%
  mutate(lab= "babylabnijmegen")
```

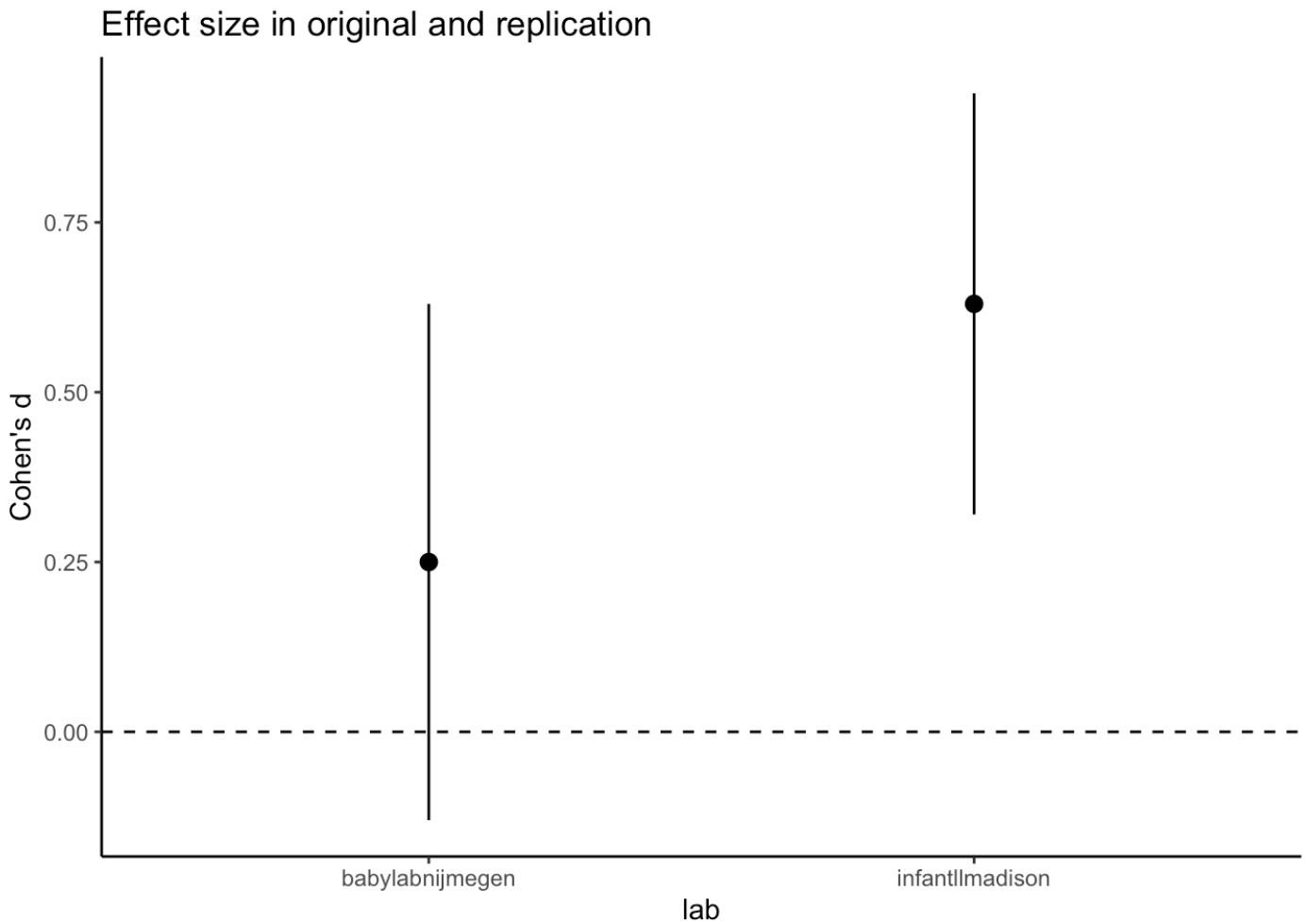
```
infantllmadison_effect_size <-
  mes(many_babies_data_summary %>% filter(lab == "infantllmadison", condition == "IDS") %>% pull(mean_prop_right), # m.1
    many_babies_data_summary %>% filter(lab == "infantllmadison", condition == "ADS") %>% pull(mean_prop_right),# m.2
    many_babies_data_summary %>% filter(lab == "infantllmadison", condition == "IDS") %>% pull(sd), #sd.1
    many_babies_data_summary %>% filter(lab == "infantllmadison", condition == "ADS") %>% pull(sd), #sd.2
    many_babies_data_summary %>% filter(lab == "infantllmadison", condition == "IDS") %>% pull(n), #n.1
    many_babies_data_summary %>% filter(lab == "infantllmadison", condition == "ADS") %>% pull(n), #n.2
    verbose = F)  %>%
  mutate(lab = "infantllmadison")
```

```
both_es <- bind_rows(infantllmadison_effect_size, babylabnijmegen_effect_size)

tidy_es <- both_es %>%
  select(lab, d, l.d, u.d) %>%
  rename(ci_lower = l.d,
         ci_upper = u.d)

kable(tidy_es)
```

lab	d	ci_lower	ci_upper
infantllmadison	0.63	0.32	0.94
babylabnijmegen	0.25	-0.13	0.63

[b]

```
ggplot(tidy_es, aes(y = d, x= lab)) +
  geom_pointrange(aes(ymin = ci_lower, ymax = ci_upper)) +
  ggtitle("Effect size in original and replication") +
  geom_hline(aes(yintercept = 0), linetype = 2) +
  ylab("Cohen's d") +
  theme_classic()
```



Question 9

This question uses the data from the Many Babies project from Assignment 5.

[a] Calculate effect sizes for the replications in two of the labs. Specifically, calculate an effect size for the replication experiment conducted by the "babylabnijmegen" lab and an effect size for the replication experiment conducted by the "infantllmadison" lab.
[b] Plot the two effect sizes.
[c] Which lab had the largest effect size? State the point estimate and confidence interval for the effect size for that lab.
[d] Interpret the effect sizes. Explain what it means to have a large effect size in this experiment. Would you guess that these two effect sizes are statistically different from each other? Why or why not?
[e] Imagine we ran another replication of this experiment and got a *negative* effect size. What would this mean?

[c]

infantllmadison had a larger effect size. The point estimate for effect size for that lab is 0.63, with a confidence interval of 0.32 to 0.94. The way this is conventionally reported in text is Cohen's d is .63 [95% CI: .32, .94]. Note that there are no units!

[d]

A larger effect size in this experiment means that the difference in looking time between Adult Directed Speech (ADS) and Infant Directed Speech (IDS) is longer. I.e., using infants look longer at the screen when infant directed speech is playing, relative using adult directed speech. We would interpret this to mean that infants have a STRONGER preference for IDS relative to ADS.

The set of "plausible values" for the means in the two conditions is very similiar (i.e. overlapping). Thus, it is unlikely that these two means are statistically different from each other.

Question 9

This question uses the data from the Many Babies project from Assignment 5.

[a] Calculate effect sizes for the replications in two of the labs. Specifically, calculate an effect size for the replication experiment conducted by the "babylabnijmegen" lab and an effect size for the replication experiment conducted by the "infantllmadison" lab.
[b] Plot the two effect sizes.
[c] Which lab had the largest effect size? State the point estimate and confidence interval for the effect size for that lab.
[d] Interpret the effect sizes. Explain what it means to have a large effect size in this experiment. Would you guess that these two effect sizes are statistically different from each other? Why or why not?
[e] Imagine we ran another replication of this experiment and got a *negative* effect size. What would this mean?

[e]
If we got a negative effect size, this would mean that the mean looking time was longer for adult directed speech than with infant directed speech (direction of difference is flipped).

Cohen's d = `(mean1 - mean2)/standard_deviation`
You can choose whether an effect is positive or negative based on how you define mean1 and mean2. In general, you want to define which is mean1 and which is mean2 so that you get a postive effect size if your hypothesis is correct. In this case, the hypothesis is "Infants prefer infant directed speech". So we'd want to define mean1 to be mean looking time in the IDS condition and mean2 to to be mean looking time in the adult directed speech condition. Defined this way, we will get a postive value when mean1 > mean2 (i.e. when infants look longer at IDS relative to ADS).

Question 10 How are the functions `filter` and `distinct` similar? How are they different? Use data from the Many Babies project to demonstrate your answer. Your answer should involve both code *and* a clear explanation.

```
many_babies_data %>%
  filter(lab== "babylabnijmegen")
```

```
## # A tibble: 110 x 5
##    lab            subid   age_days condition mean_looking_time
##    <chr>          <chr>      <dbl> <chr>                <dbl>
##  1 babylabnijmegen ba0169       245 ADS                   8.38
##  2 babylabnijmegen ba0169       245 IDS                   7.28
##  3 babylabnijmegen ba01912      323 ADS                   5.28
##  4 babylabnijmegen ba01912      323 IDS                   5.21
##  5 babylabnijmegen ba0269       206 ADS                   5.20
##  6 babylabnijmegen ba0269       206 IDS                   9.33
##  7 babylabnijmegen ba02912      335 ADS                   9.32
##  8 babylabnijmegen ba02912      335 IDS                   9.43
##  9 babylabnijmegen ba0369       257 ADS                   8.42
## 10 babylabnijmegen ba0369       257 IDS                   7.35
## # … with 100 more rows
```

```
many_babies_data %>%
  distinct(lab)
```

```
## # A tibble: 6 x 1
##   lab
##   <chr>
## 1 babylabnijmegen
## 2 babylabparisdescartes1
## 3 babylabplymouth
## 4 babylabpotsdam
## 5 babylabprinceton
## 6 infantllmadison
```

Question 11 Load the data from Study 1 into R. Sort the data so that children who have the highest value for `stereo` are at the top. Show the first 7 rows of this dataset.

```
lian_data <- read_csv('data/_Study1.csv')
```

```
lian_data %>%
  arrange(-stereo) %>%
  slice(1:7) %>%
  kable()
```

subj	gender	age	trait	stereo
5	1	7	1	1
6	1	5	1	1
22	2	5	1	1
23	1	5	1	1
88	1	7	1	1
90	1	7	1	1
94	1	5	1	1

32 children participated in each age group.

Question 12 How many children participated in the experiment in each age group?

```
lian_data_tidy %>%
  distinct(subj, .keep_all = T) %>%
  count(age) %>%
  kable()
```

age	n
5	32
6	32
7	32

Question 13 Calculate the standard error of the mean for the dependent variable (mean proportion of times children linked the target trait to their own gender) for girls on the “nice” trait task.

```
lian_data_tidy %>%
  filter(gender == "girl", trait == "nice") %>%
  summarize(n = n(),
            sd = sd(stereo)) %>%
  mutate(sem = sd/sqrt(n)) %>%
  kable()
```

n	sd	sem
48	0.2200663	0.0317638

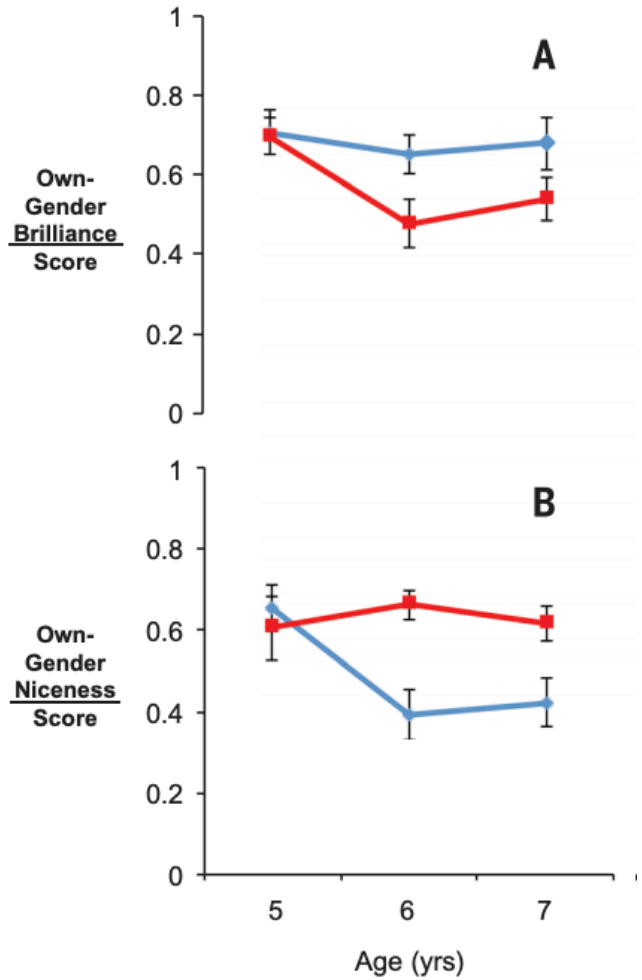
Question 14 Recreate Figure 1A and Figure 1B from the paper. Include error bars that are 95% confidence intervals (rather than standard errors).

REPORT

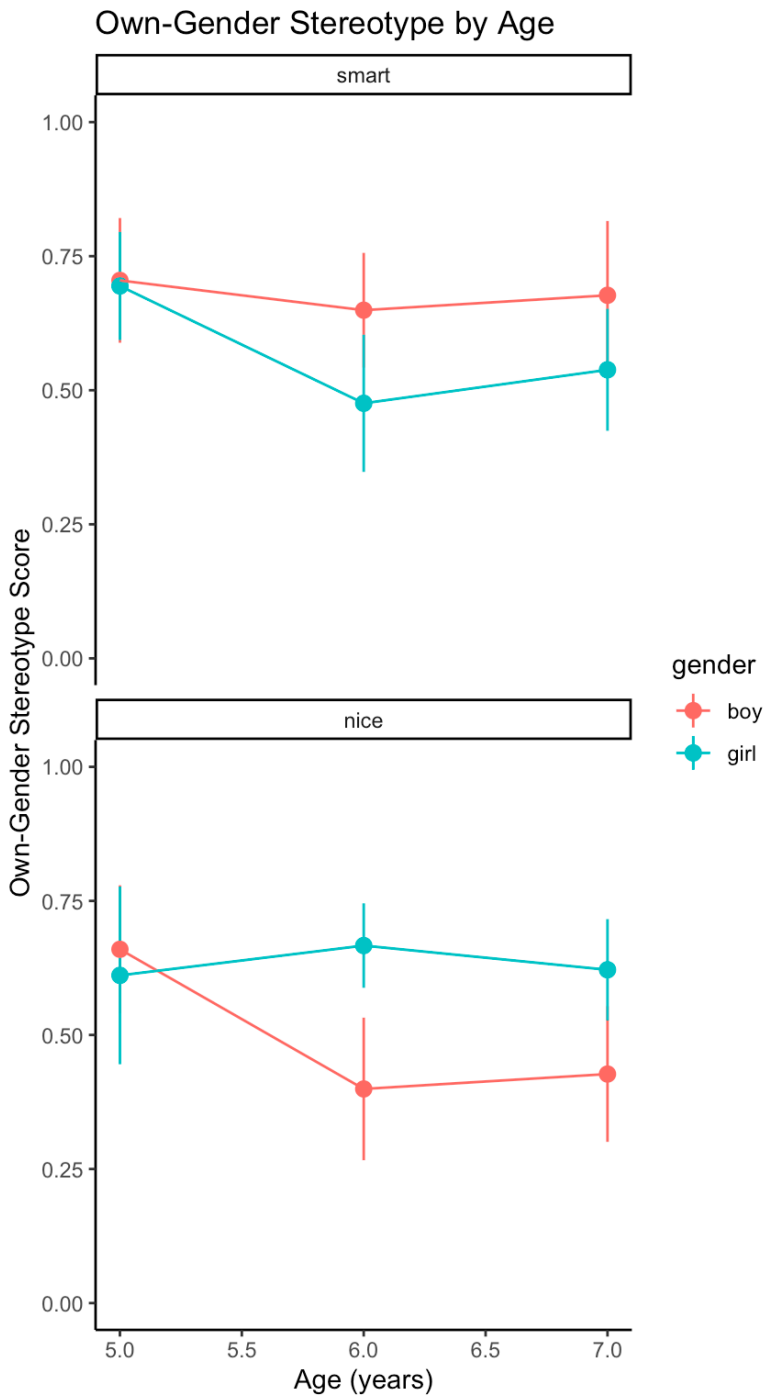
PSYCHOLOGY

Gender stereotypes about intellectual ability emerge early and influence children's interests

Lin Bian,[1,2]* Sarah-Jane Leslie,[3] Andrei Cimpian[1,2]*



```r
plotting_data <- lian_data_tidy %>%
  mutate(trait = fct_rev(trait)) %>%
  group_by(trait, gender, age) %>%
  summarize(mean = mean(stereo),
            sd = sd(stereo),
            n = n()) %>%
  mutate(ci_range_95 =  qt(1 - (0.05 / 2), n - 1) * (sd/sqrt(n)),
         ci_lower = mean - ci_range_95,
         ci_upper = mean + ci_range_95)

ggplot(plotting_data, aes(x = age, y = mean, color = gender))  +
  geom_pointrange(aes(ymin = ci_lower, ymax = ci_upper)) +
  ylim(0,1) +
  facet_wrap(~trait, ncol = 1) +
  geom_line() +
  ylab("Own-Gender Stereotype Score") +
  xlab("Age (years)") +
  ggtitle("Own-Gender Stereotype by Age") +
  theme_classic()
```



Question 15 Look at Figure 1 in the paper. Compare the error bars for the seven-year-old girls in panel B (Study 1) and panel D (Study 2). State one reason why they might be smaller in Study 2 compared to Study 1.

The means might be smaller in Study 2, compared to Study 1, because there were more participants in Study 2 (N = 24) than in study 1 (N = 16). The size of the a confidence interval depends on standarder error, and standard error depends on sample size:

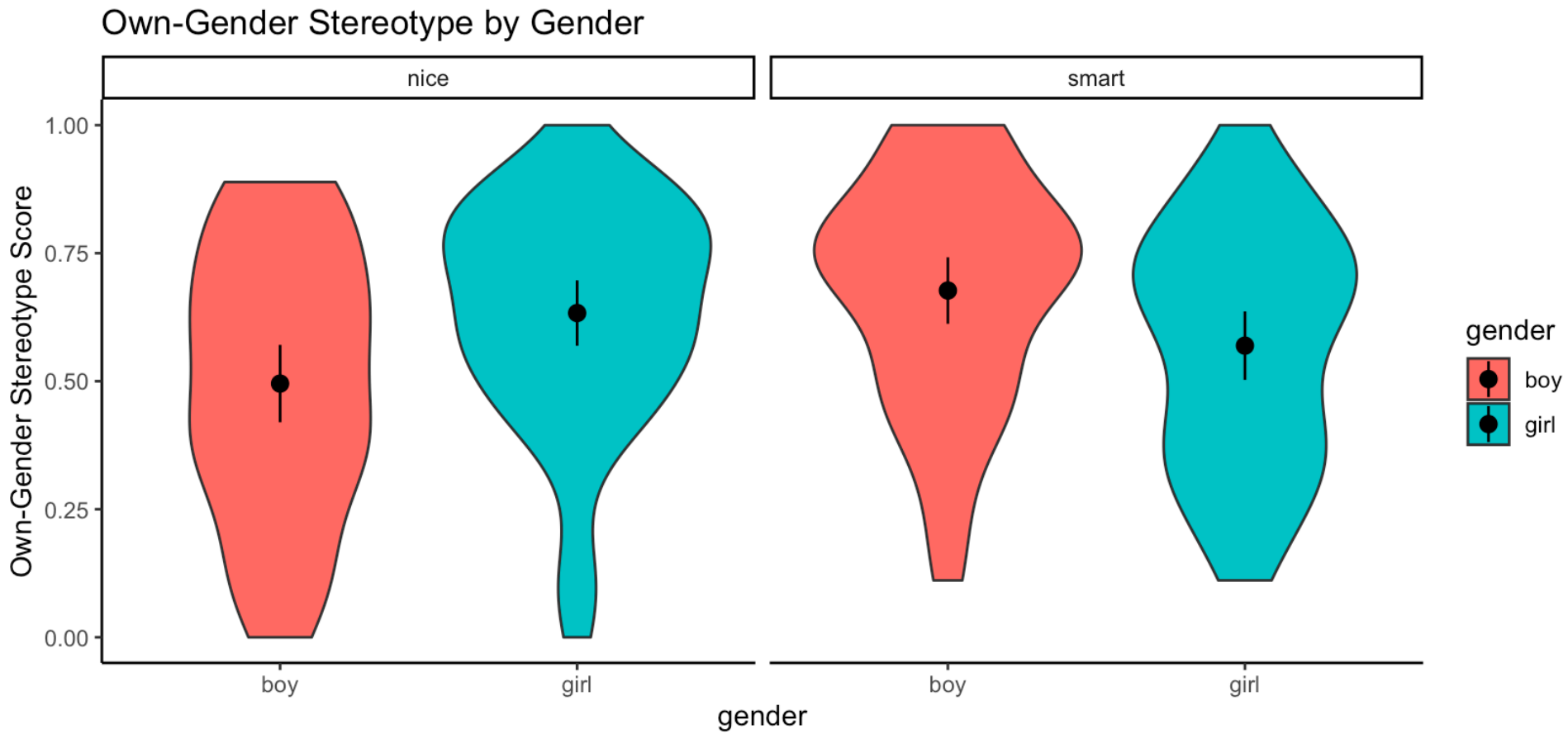
95% CI width = $\sim 1.96 * SE$

SE = $sd/sqrt(n)$.

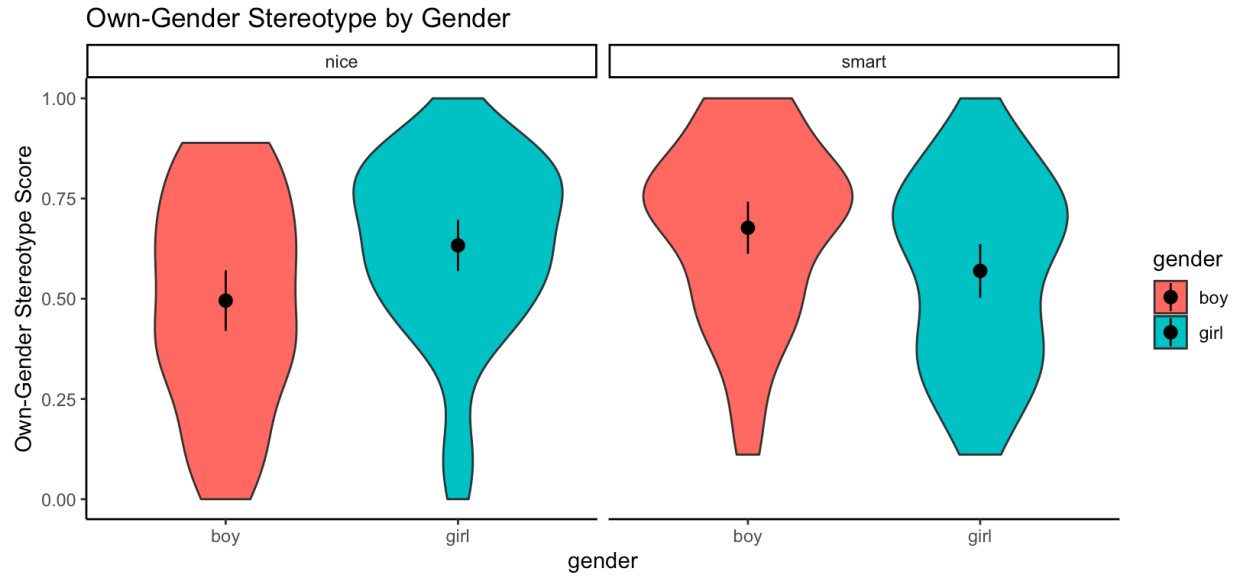
Larger sample size means smaller SE and therefore smaller CI (i.e. smaller range of plausible values for the mean).

Question 16

Extra credit: Use the data from Study 1 to make a second plot. Make a violin plot (`geom_violin`) showing the distribution of the mean proportion of times children linked the target trait to their own gender. Show a separate violin for each unique combination of gender and trait. Add a point estimate and a confidence interval to each violin showing the corresponding 95% confidence interval.

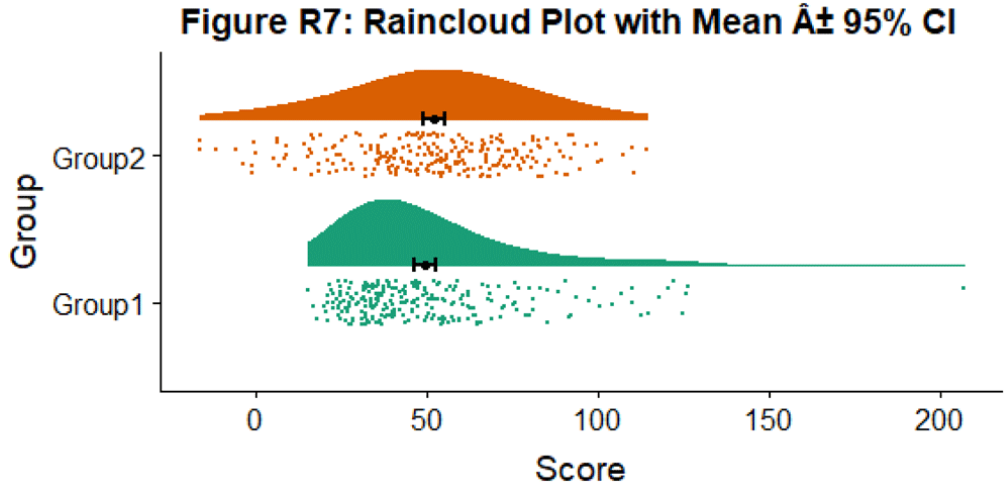


```
plotting_data2 <- lian_data_tidy %>%
  group_by(trait, gender) %>%
  summarize(mean = mean(stereo),
            sd = sd(stereo),
            n = n()) %>%
  mutate(ci_range_95 =  qt(1 - (0.05 / 2), n - 1) * (sd/sqrt(n)),
         ci_lower = mean - ci_range_95,
         ci_upper = mean + ci_range_95)

ggplot(data = lian_data_tidy)  +
  geom_violin(data = lian_data_tidy, aes(x = gender, y = stereo, fill = gender)) +
  geom_pointrange(data = plotting_data2,
                  aes(x = gender, y = mean, fill = gender, ymin = ci_lower, ymax = ci_upper)) +
  ylim(0,1) +
  facet_wrap(~trait) +
  ylab("Own-Gender Stereotype Score") +
  ggtitle("Own-Gender Stereotype by Gender") +
  theme_classic()
```



What are violin plots?

- Same as probability density (geom_density()), except mirrored/doubled and turned on its side
- Often preferable to bar graphs because it allows you to see the distribution of the underlying data
- Another alternative "rain cloud" plots



On Friday

- (will post solutions to midterm online)
- More practice with effect sizes
- Reading:

Chapter 18 Quantifying effects and designing studies

In the previous chapter we discussed how we can use data to test hypotheses. Those methods provided a binary answer: we either reject or fail to reject the null hypothesis. However, this kind of decision overlooks a couple of important questions. First, we would like to know how much uncertainty we have about the answer (regardless of which way it goes). In addition, sometimes we don't have a clear null hypothesis, so we would like to see what range of estimates are consistent with the data. Second, we would like to know how large the effect actually is, since as we saw in the weight loss example in the previous chapter, a statistically significant effect is not necessarily a practically important effect.

In this chapter we will discuss methods to address these two questions: confidence intervals to provide a measure of our uncertainty about our estimates, and effect sizes to provide a standardized way to understand how large the effects are. We will also discuss the concept of *statistical power* which tells us how well we can expect to find any true effects that might exist.