Statistical Foundations: Effect Sizes

11 October 2020

Modern Research Methods

Midterm

- Handed out next Tuesday 19th at noon, due Thursday 21st at noon
- Open book, but must complete on your own
- Similar to assignments, but longer and I will give you less code/structure
- I'll provide one or more datasets and you'll have to analyze/plot it
- Also, conceptual questions
- Will cover all material through next Monday
- Lab this Friday will be review – I won't prepare anything, I'll ask you to come prepared with specific topics you'd like to review

The goal of an experiment is to estimate population values (e.g., means).

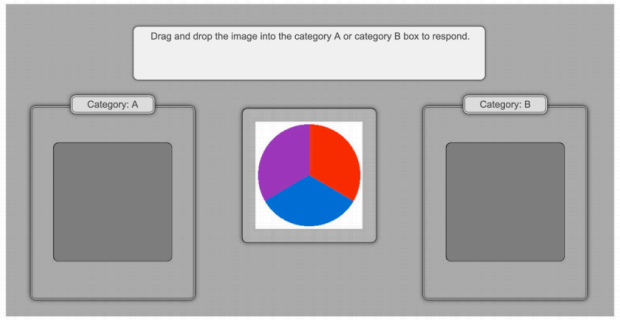
- But, can only observe sample of population in each experiment.
- Use sample mean to estimate population mean.
- We expect our estimation to not be perfect.
- If our estimation is roughly right, and we run the experiment again ("replicate it"), should get roughly the **same value**.

Sampling Distribution

one experiment .84

Replicating Zettersten and Lupyan (2020)

Original

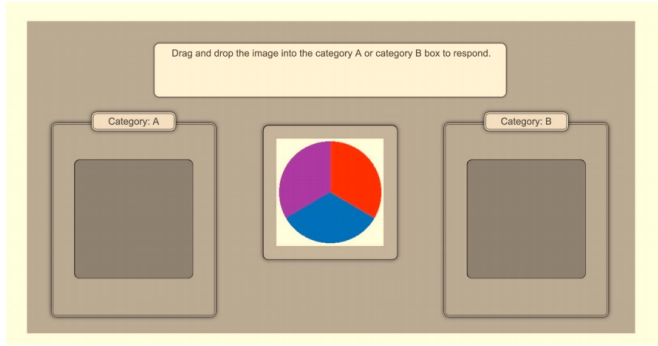


CLAIM: It's easier to learn a category when the colors are nameable.

predicting participants' trial-by-trial accuracy on training trials from condition, including a by-subject random intercept.[3] We used the lme4 package version 1.1-21 in R (version 3.6.1) to fit all models (D. Bates & Maechler, 2009; R Development Core Team, 2019). Participants in the High Nameability condition (M = 84.0%, 95% CI = [78.6%, 89.4%]) were more accurate than participants in the Low Nameability Condition (M = 67.7%, 95% CI = [59.9%, 75.4%]), b = 1.02, 95% Wald

Replication

[Us]



High Nameability Condition = 75%
Low Nameability Condition = 69%

Did we replicate it?

Confidence intervals give plausible range of values for population estimate

Calculate a confidence intervals for the means on accuracy reported in Zettersten and Lupyan (2020), Experiment 1A

```
DATA_PATH <- "https://osf.io/a4dzb/download"
zl_data <- read_csv(DATA_PATH)
```

```
zl_clean <- zl_data %>%
     clean_names() %>%
     select(experiment, subject, age, condition, block_num, is_right)
```

```
zl_exp1a <- zl_clean %>%
     filter(experiment == "1A")
```

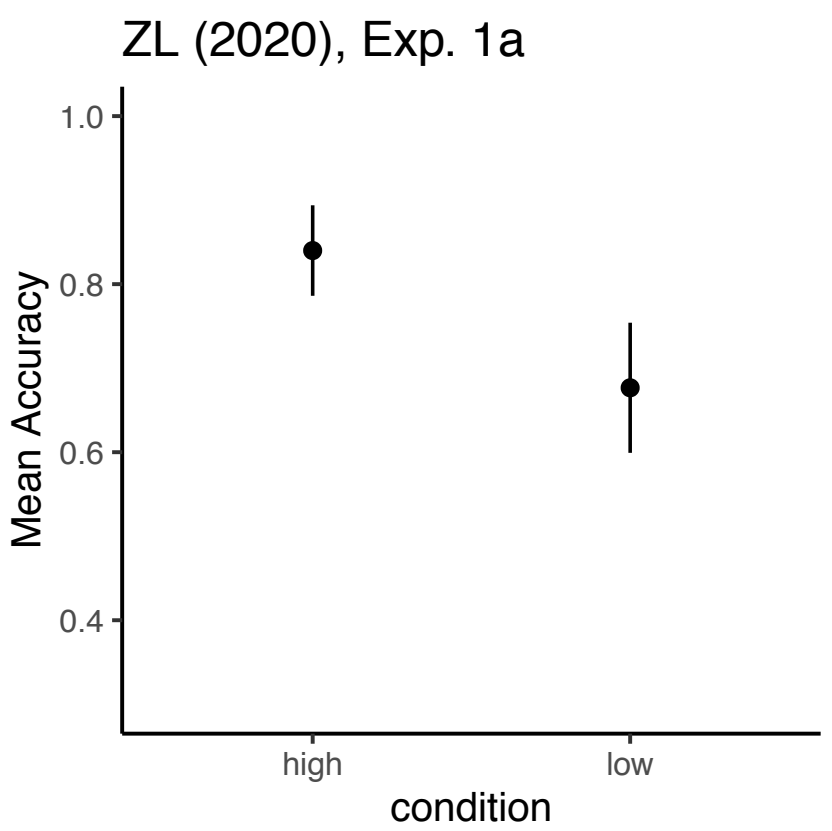
Review: Calculating and plotting CIs

```
ms_by_overall <- zl_exp1a %>%
  group_by(subject, condition) %>%
  summarize(prop_right = sum(is_right)/n())
```

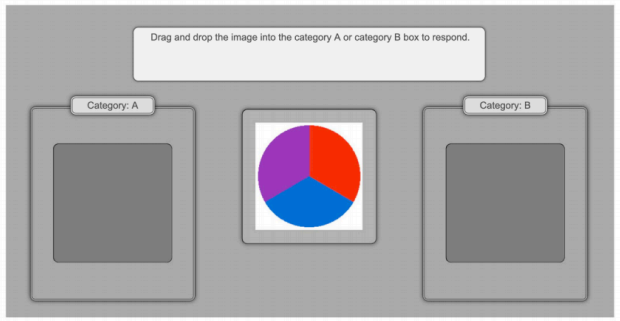
```
means_by_condition_with_ci_t <- ms_by_overall %>%
  group_by(condition) %>%
  summarize(mean = mean(prop_right),
            sd = sd(prop_right),
            n = n()) %>%
  mutate(ci_range_95 =  qt(1 - (0.05 / 2), n - 1) * (sd/sqrt(n)),
         ci_lower = mean - ci_range_95,
         ci_upper = mean + ci_range_95)
```

Review: Calculating and plotting CIs

```
ggplot(means_by_condition_with_ci_t, aes(x = condition, y = mean)) +
  geom_pointrange(aes(ymin = ci_lower, ymax = ci_upper), size = 1) +
  ylim(.3, 1) +
  ylab("Mean Accuracy") +
  ggtitle("ZL (2020), Exp. 1a")  +
  theme_classic(base_size = 24)
```

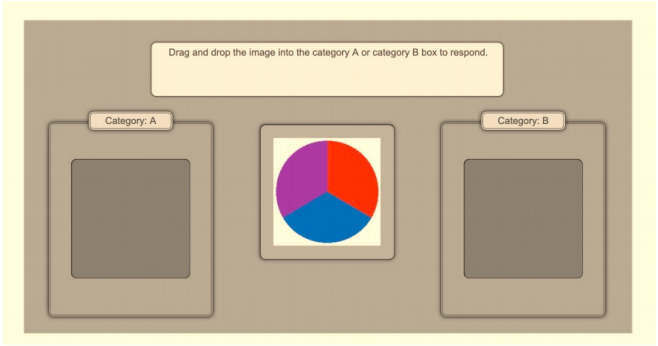


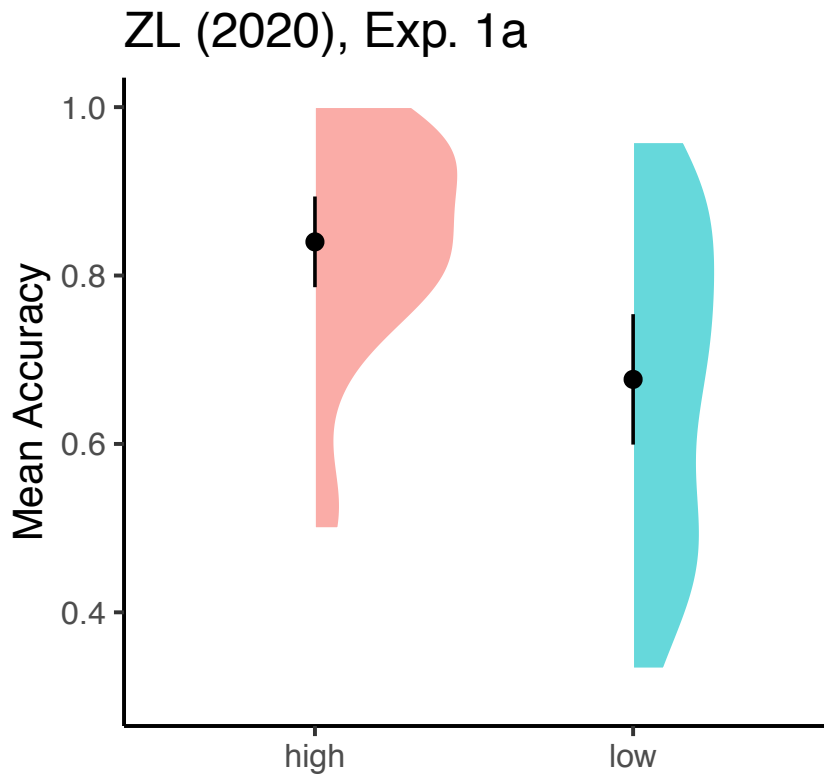
Original

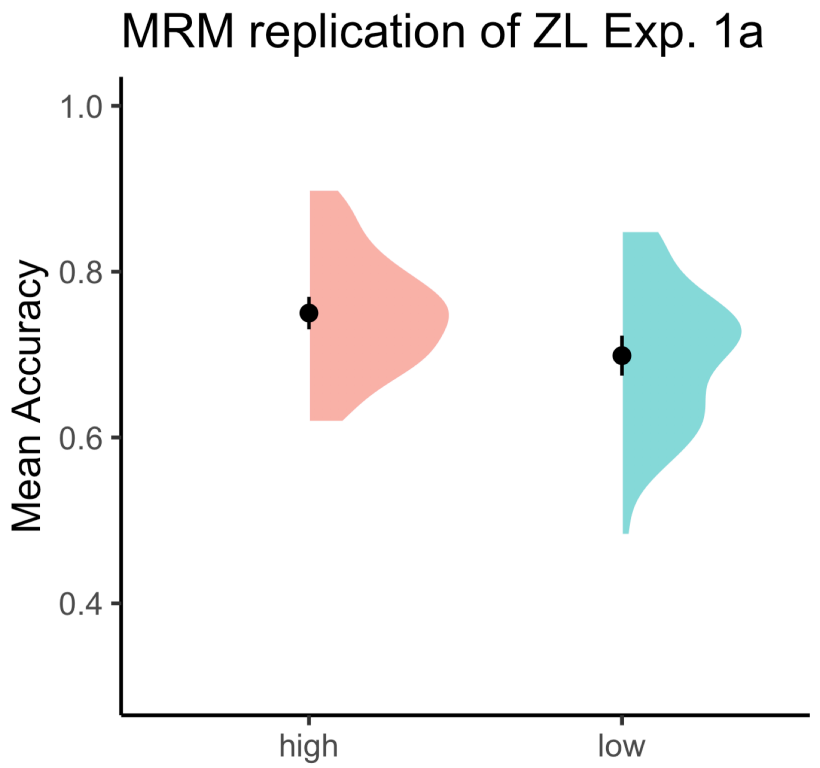


Replication

[Us]

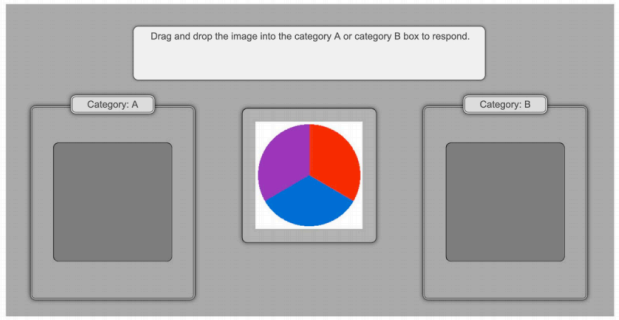






Replicating Zettersten and Lupyan (2020)

Original

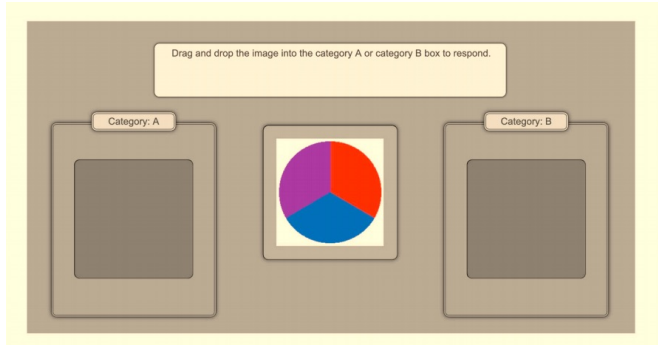


CLAIM: It's easier to learn a category when the colors are nameable.

predicting participants' trial-by-trial accuracy on training trials from condition, including a by-subject random intercept.[3] We used the lme4 package version 1.1-21 in R (version 3.6.1) to fit all models (D. Bates & Maechler, 2009; R Development Core Team, 2019). Participants in the High Nameability condition (M = 84.0%, 95% CI = [78.6%, 89.4%]) were more accurate than participants in the Low Nameability Condition (M = 67.7%, 95% CI = [59.9%, 75.4%]), b = 1.02, 95% Wald

Replication

[Us]

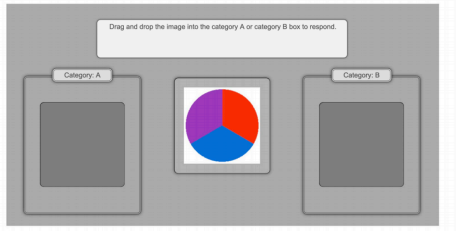


High Nameability Condition = 75%
Low Nameability Condition = 69%

Did we replicate it? YES!

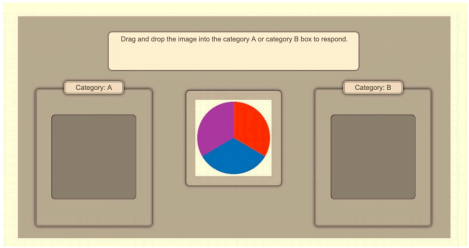
In which experiment is the effect bigger?

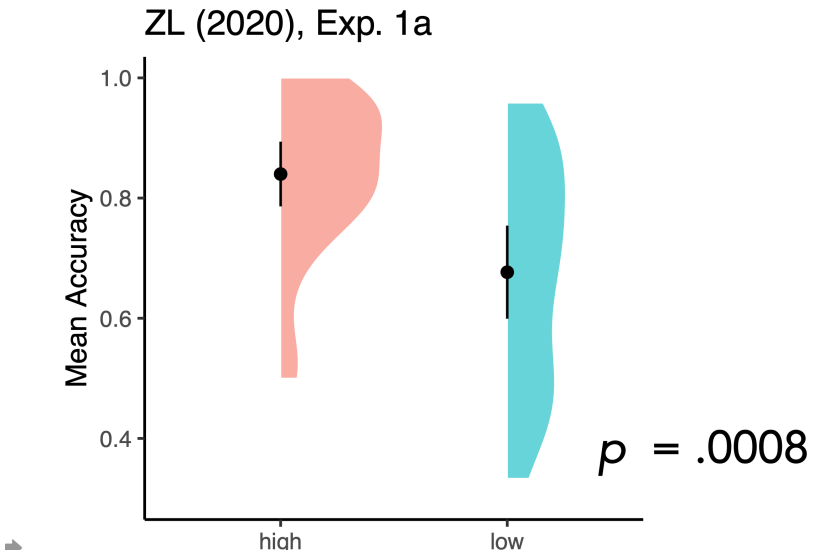
Original

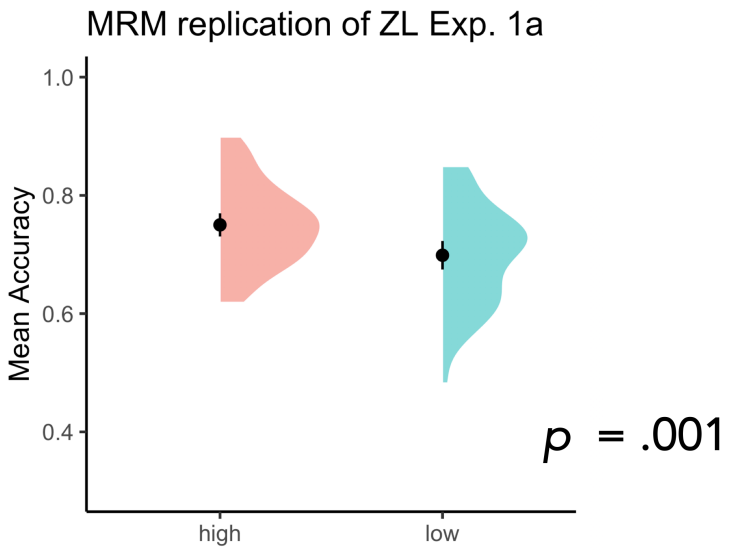


Replication

[Us]







Quantifying the magnitude of an effect

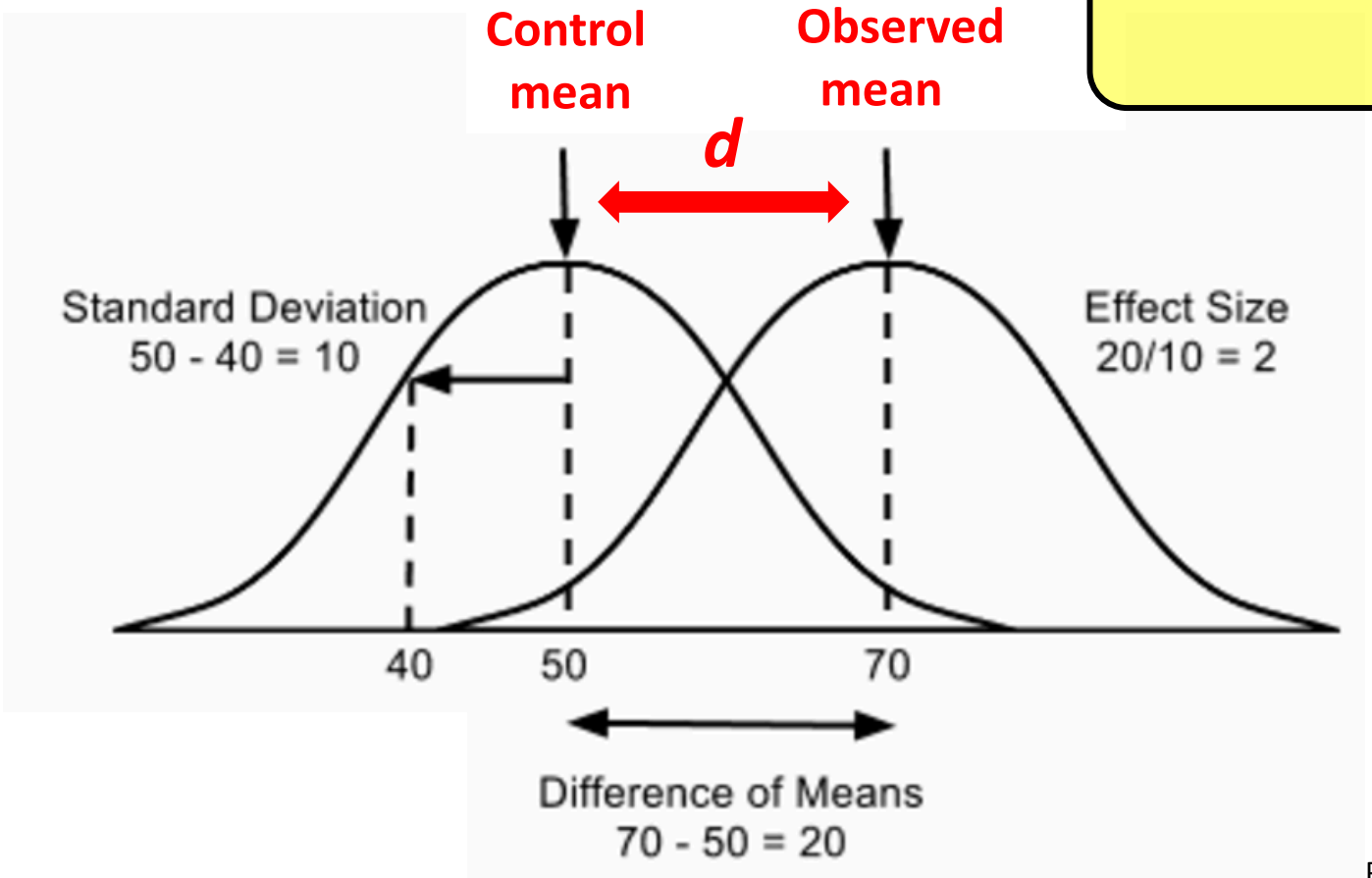
Effect size as unit of analysis: Quantitative, scale-free measure of an effect.

Cohen's *d*:

$$\textbf{Effect Size} = \frac{\textbf{diff. between means}}{\textbf{standard dev.}}$$

Quantifying the magnitude of an effect

$$\text{Effect Size} = \frac{\text{diff. between means}}{\text{standard dev.}}$$



Plot from: http://www.realeducationreform.com/questions.html

Explore Cohen's *d*

https://rpsychologist.com/d3/cohend/

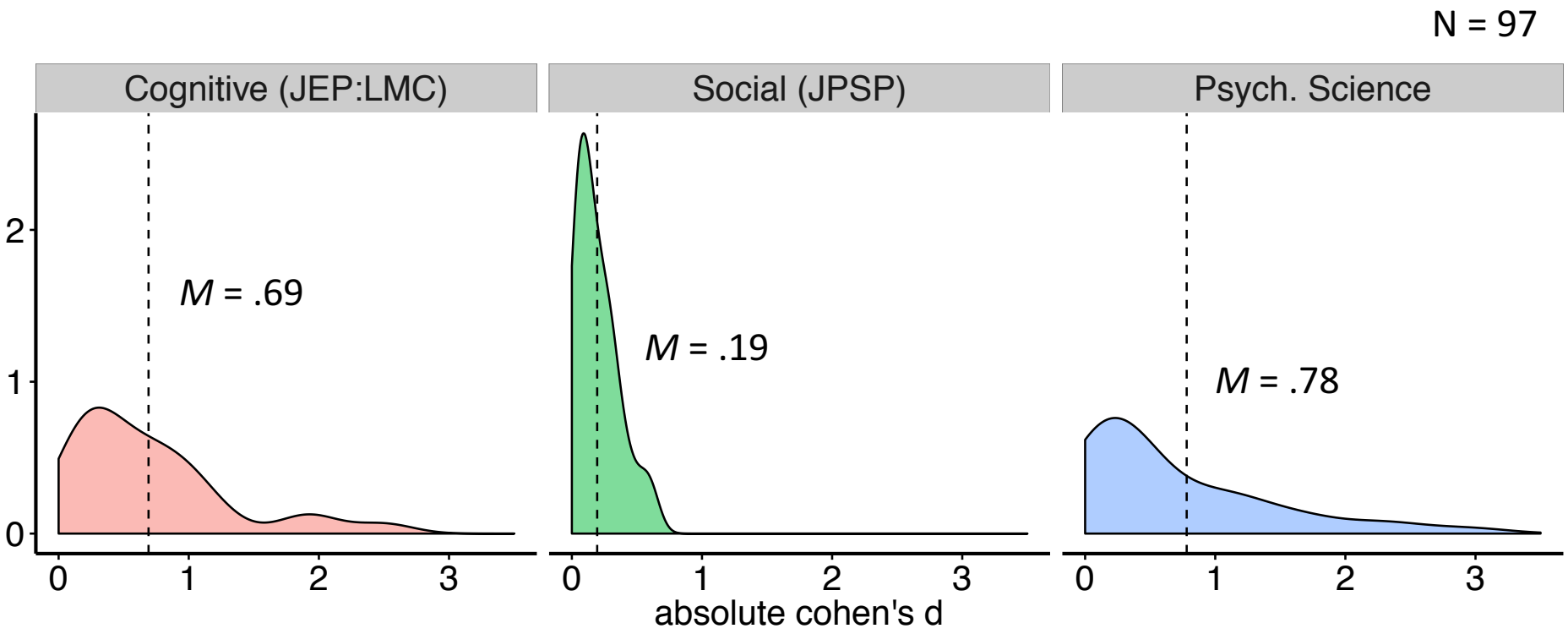
Interpreting Cohen's *d*

Size	Description	Cohen's Intuition	Psychological Example
.2	"small"	Diff. between the heights of 15 yo and 16 yo girls in the US	Bouba-kiki effect in kids (~.15; Lammertink, et al. 2016)
.5	"medium"	Diff. between the heights of 14 yo and 18 yo girls.	Cognitive behavioral therapy on anxiety (~ .4; (Belleville, et al., 2004) Sex difference in implicit math attitudes (~.5; Klein, et al., 2013)
.8	"large"	Diff. between the heights of 13 yo and 18 yo girls.	Syntactic Priming (~.9; Mahowald, et al., under review) Mutual exclusivity (~1.0; Lewis & Frank, in prep)

(Cohen, 1969)

Interpreting Cohen's *d*

Estimating the Replicability of Psychological Science (OSF, *Science*, 2015)



Relatively "large" effects reported in cognitive psychology

Effect sizes

Prototype: Cohen's *d*

- Depends on aspect of design (e.g., within vs. between subject)
- Many effect size metrics (Hedge's *g* for small samples)
- Can convert between ES metrics
- Can calculate via different pieces of raw data

Multiple Types:

- the difference is between groups (t-test, *d*)
- the relationship between variables (correlation, *r)*
- the amount of variance accounted for by a factor (ANOVA, regression, *f)*
- More generally, for any statistical test you conduct, can compute effect size (some more straight-forwardly than others)

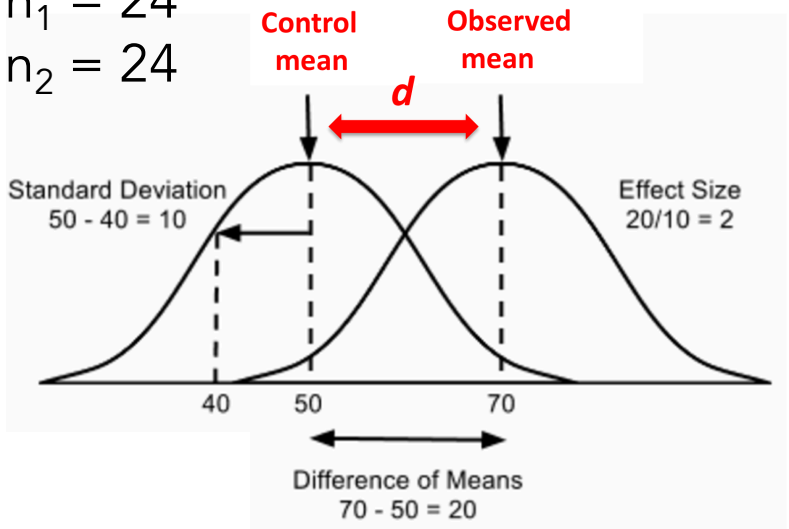
TABLE 1 **Common Effect Size Indices**[a]

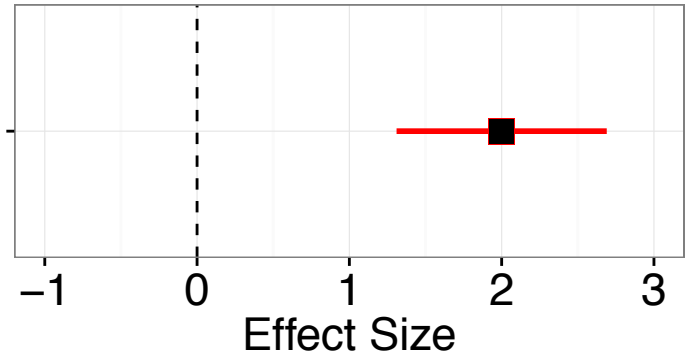
Index	Description[b]	Effect Size	Comments
Between groups			
Cohen's *d*[a]	$d = M_1 - M_2 / s$ $M_1 - M_2$ is the difference between the group means (M); s is the standard deviation of either group	Small 0.2 Medium 0.5 Large 0.8 Very large 1.3	Can be used at planning stage to find the sample size required for sufficient power for your study
Odds ratio (OR)	$\frac{\text{Group 1 odds of outcome}}{\text{Group 2 odds of outcome}}$ If OR = 1, the odds of outcome are equally likely in both groups	Small 1.5 Medium 2 Large 3	For binary outcome variables Compares odds of outcome occurring from one intervention vs another
Relative risk or risk ratio (RR)	Ratio of probability of outcome in group 1 vs group 2; If RR = 1, the outcome is equally probable in both groups	Small 2 Medium 3 Large 4	Compares probabilities of outcome occurring from one intervention to another
Measures of association			
Pearson's *r* correlation	Range, −1 to 1	Small ±0.2 Medium ±0.5 Large ± 0.8	Measures the degree of linear relationship between two quantitative variables
r^2 coefficient of determination	Range, 0 to 1; Usually expressed as percent	Small 0.04 Medium 0.25 Large 0.64	Proportion of variance in one variable explained by the other

(Sullivan & Feinn, 2012)

Effect size confidence interval

$n_1 = 24$
$n_2 = 24$

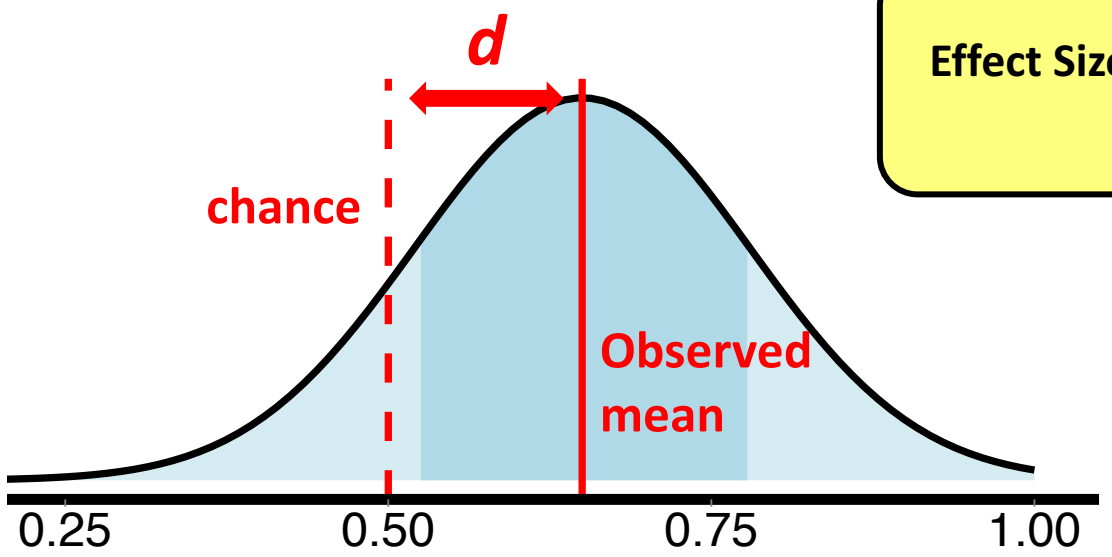




$$
\begin{aligned}
var_d &= \frac{n_1 + n_2}{n_1 * n_2} + \frac{d^2}{2(n_1 + n_2)} \\
&= \frac{24 + 24}{24 * 24} + \frac{2^2}{2(24 + 24)} \\
&= .125
\end{aligned}
$$

$$
\begin{aligned}
CI(d) &= Est(d) \pm z_{(\alpha/2)} * \sqrt{var(d)} \\
&= 2 \pm 1.96 * .35 \\
&= 2 \pm .69
\end{aligned}
$$

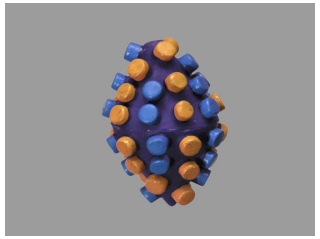
The one-sample case



Effect Size = $\frac{\text{diff. between means}}{\text{standard dev.}}$

Example: Mutual exclusivity effect size

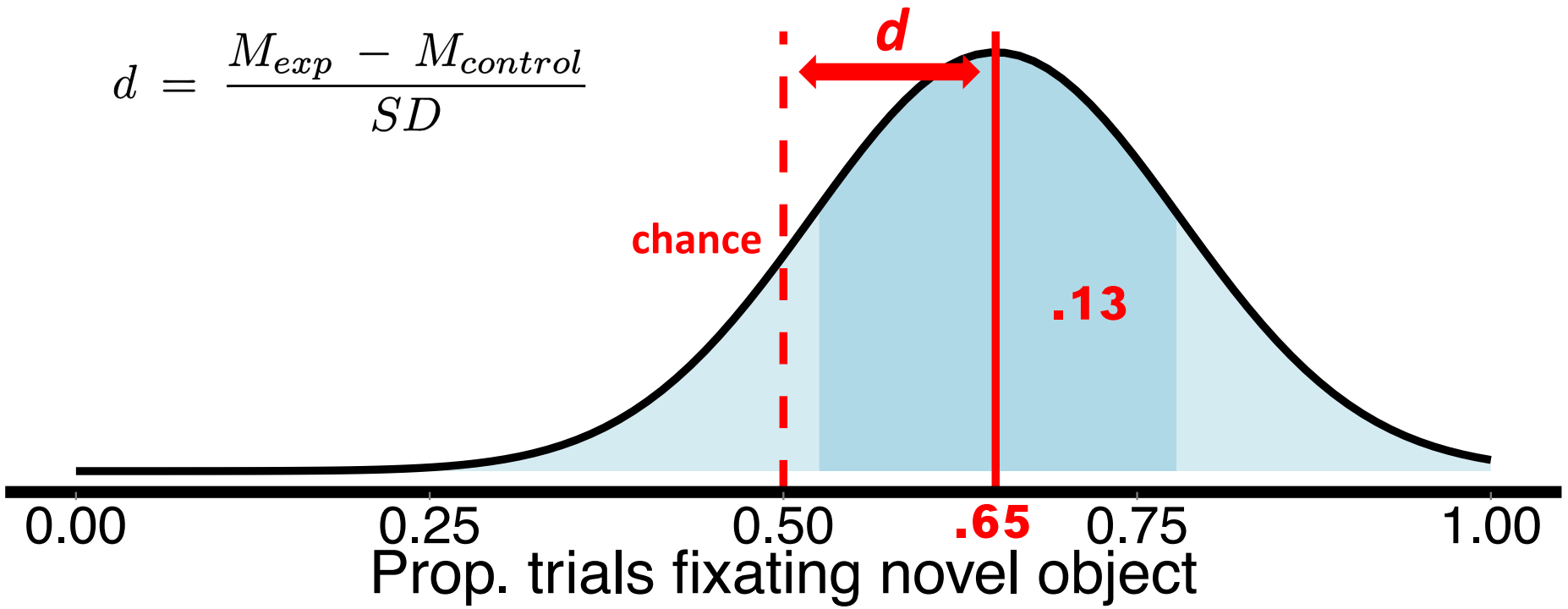
Where's the dofa?



Bion, et al. (2013)

For 24 mo, mean proportion of trials fixating on novel object = .65 (SD = .13)

$$d = \frac{M_{exp} - M_{control}}{SD}$$



We're going to calculate a confidence intervals for the means on accuracy reported in Zettersten and Lupyan (2020), Experiment 1A

Let's start by loading the data.

```
DATA_PATH <- "https://osf.io/a4dzb/download"

zl_data <- read_csv(DATA_PATH)

zl_clean <- zl_data %>%
     clean_names() %>%
     select(experiment, subject, age, condition, block_num, is_right)

zl_exp1a <- zl_clean %>%
     filter(experiment == "1A")

ms_by_overall<- zl_exp1a %>%
  group_by(subject, condition) %>%
  summarize(prop_right = sum(is_right)/n())  %>%
   mutate(experiment = "original_ZL2020")

ms_by_overall_replication <- read_csv("mrm_replication_data.csv") %>%
  rename(subject = subid) %>%
  mutate(subject = as.character(subject))

all_data <- bind_rows(ms_by_overall, ms_by_overall_replication) %>%
  select(experiment, subject, everything()) %>%
  ungroup()
```

The data:

experiment	subject	condition	prop_right
original_ZL2020	p150212	low	0.8750000
original_ZL2020	p157080	low	0.7083333
original_ZL2020	p191463	low	0.9583333
original_ZL2020	p20905	high	0.9583333
original_ZL2020	p213384	high	1.0000000

```
all_data %>%
  group_by(experiment, condition) %>%
  summarize(n = n())
```

```
## # A tibble: 4 × 3
## # Groups:   experiment [2]
##   experiment                condition     n
##   <chr>                     <chr>     <int>
## 1 MRM_replication_of_LZ2020 high         50
## 2 MRM_replication_of_LZ2020 low          50
## 3 original_ZL2020           high         25
## 4 original_ZL2020           low          25
```

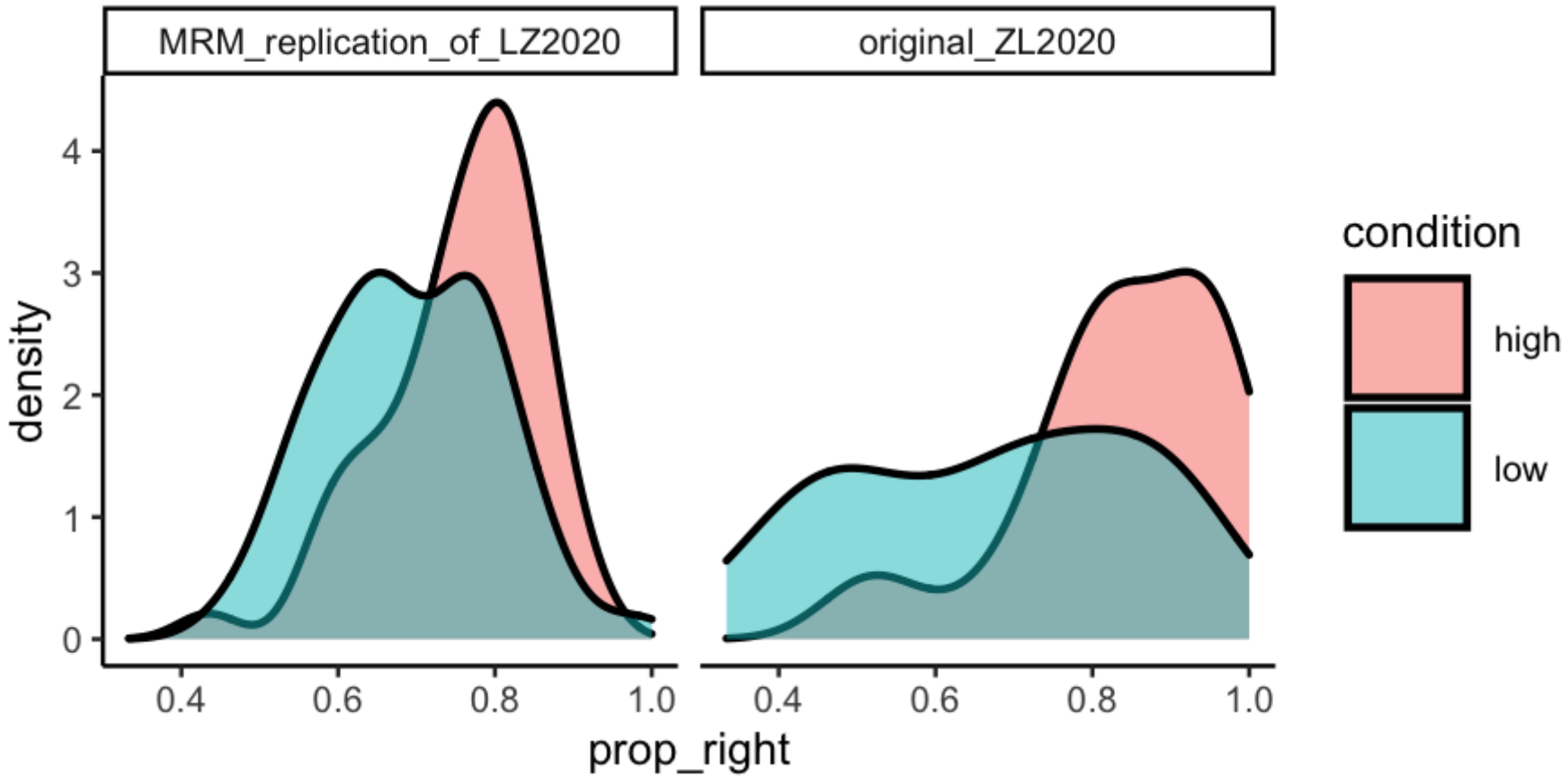
--

Or, equivalently:

```
all_data %>%
  count(experiment, condition)
```

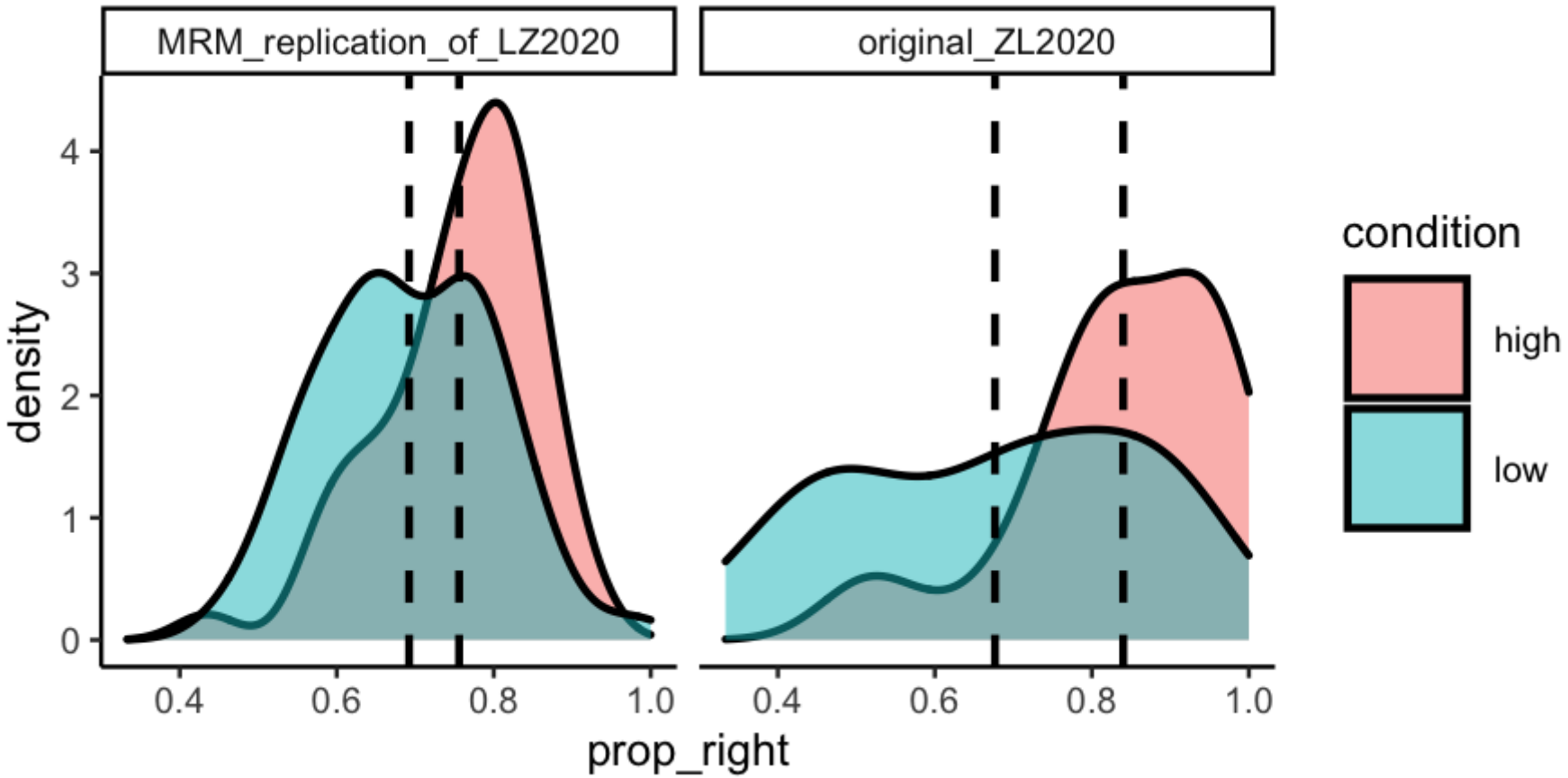
```
## # A tibble: 4 × 3
##   experiment                condition     n
##   <chr>                     <chr>     <int>
## 1 MRM_replication_of_LZ2020 high         50
## 2 MRM_replication_of_LZ2020 low          50
## 3 original_ZL2020           high         25
## 4 original_ZL2020           low          25
```

```
ggplot(all_data, aes(x = prop_right, fill = condition)) +
  geom_density(alpha = .5) +
  facet_wrap(~experiment)
```



experiment	condition	mean_prop_right	sd	n
MRM_replication_of_LZ2020	high	0.7558991	0.0989817	50
MRM_replication_of_LZ2020	low	0.6922030	0.1099081	50
original_ZL2020	high	0.8400000	0.1304817	25
original_ZL2020	low	0.6766667	0.1876080	25

```
ggplot(all_data, aes(x = prop_right, fill = condition)) +
  geom_density(alpha = .5) +
  geom_vline(data = dist_summary, aes(xintercept = mean_prop_right), linetype = 2) +
  facet_wrap(~experiment)
```



experiment	condition	mean_prop_right	sd	n
MRM_replication_of_LZ2020	high	0.7558991	0.0989817	50
MRM_replication_of_LZ2020	low	0.6922030	0.1099081	50
original_ZL2020	high	0.8400000	0.1304817	25
original_ZL2020	low	0.6766667	0.1876080	25

```
dist_summary %>%
  filter(experiment == "MRM_replication_of_LZ2020", condition == "high") %>%
  pull(mean_prop_right)
```

```
## [1] 0.7558991
```

Calculate Cohen's d for a single experiment using `mes` function from `compute.es` package:

```
replication_effect_size <-
  mes(dist_summary %>% filter(experiment == "MRM_replication_of_LZ2020",
                              condition == "high") %>% pull(mean_prop_right), # m.1
    dist_summary %>% filter(experiment == "MRM_replication_of_LZ2020",
                            condition == "low") %>% pull(mean_prop_right),# m.2
    dist_summary %>% filter(experiment == "MRM_replication_of_LZ2020",
                            condition == "high") %>% pull(sd), #sd.1
    dist_summary %>% filter(experiment == "MRM_replication_of_LZ2020",
                            condition == "low") %>% pull(sd), #sd.2
    dist_summary %>% filter(experiment == "MRM_replication_of_LZ2020",
                            condition == "high") %>% pull(n), #n.1
    dist_summary %>% filter(experiment == "MRM_replication_of_LZ2020",
                            condition == "low") %>% pull(n), #n.2
    verbose = F) %>%
  mutate(experiment = "MRM_replication_of_LZ2020")

replication_effect_size
```

```
##   N.total n.1 n.2    d var.d  l.d  u.d  U3.d  cl.d cliffs.d pval.d   g var.g
## 1     100  50  50 0.61  0.04 0.21 1.01 72.87 66.66     0.33      0 0.6  0.04
##    l.g u.g  U3.g  cl.g pval.g    r var.r l.r  u.r pval.r fisher.z var.z l.z u.z
## 1 0.21   1 72.72 66.54      0 0.29  0.01 0.1 0.46      0      0.3  0.01 0.1 0.5
##     OR l.or u.or pval.or lOR l.lor u.lor pval.lor  NNT
## 1 3.02 1.46 6.25       0 1.1  0.38  1.83        0 4.81
##                   experiment
## 1 MRM_replication_of_LZ2020
```

```r
original_effect_size <-
  mes(dist_summary %>% filter(experiment == "original_ZL2020",
                              condition == "high") %>% pull(mean_prop_right),
    dist_summary %>% filter(experiment == "original_ZL2020",
                            condition == "low") %>% pull(mean_prop_right),
    dist_summary %>% filter(experiment == "original_ZL2020",
                            condition == "high") %>% pull(sd),
    dist_summary %>% filter(experiment == "original_ZL2020",
                            condition == "low") %>% pull(sd),
    dist_summary %>% filter(experiment == "original_ZL2020",
                            condition == "high") %>% pull(n),
    dist_summary %>% filter(experiment == "original_ZL2020",
                            condition == "low") %>% pull(n),
    verbose = F) %>%
  mutate(experiment = "original_ZL2020")

original_effect_size
```

```
##   N.total n.1 n.2    d var.d  l.d u.d  U3.d  cl.d cliffs.d pval.d    g var.g
## 1      50  25  25 1.01  0.09 0.42 1.6 84.39 76.26     0.53      0 0.99  0.09
##    l.g  u.g  U3.g  cl.g pval.g    r var.r  l.r  u.r pval.r fisher.z var.z  l.z
## 1 0.42 1.57 84.01 75.91      0 0.46  0.01 0.21 0.65      0      0.5  0.02 0.21
##    u.z   OR l.or u.or pval.or  lOR l.lor u.lor pval.lor  NNT      experiment
## 1 0.78 6.25 2.15 18.2       0 1.83  0.77   2.9        0 2.72 original_ZL2020
```

```
both_es <- bind_rows(original_effect_size, replication_effect_size)

both_es
```

```
##   N.total n.1 n.2    d var.d  l.d  u.d  U3.d  cl.d cliffs.d pval.d    g var.g
## 1      50  25  25 1.01  0.09 0.42 1.60 84.39 76.26     0.53      0 0.99  0.09
## 2     100  50  50 0.61  0.04 0.21 1.01 72.87 66.66     0.33      0 0.60  0.04
##    l.g  u.g  U3.g  cl.g pval.g    r var.r  l.r  u.r pval.r fisher.z var.z  l.z
## 1 0.42 1.57 84.01 75.91      0 0.46  0.01 0.21 0.65      0      0.5  0.02 0.21
## 2 0.21 1.00 72.72 66.54      0 0.29  0.01 0.10 0.46      0      0.3  0.01 0.10
##    u.z   OR l.or  u.or pval.or  lOR l.lor u.lor pval.lor  NNT
## 1 0.78 6.25 2.15 18.20       0 1.83  0.77  2.90        0 2.72
## 2 0.50 3.02 1.46  6.25       0 1.10  0.38  1.83        0 4.81
##                  experiment
## 1           original_ZL2020
## 2 MRM_replication_of_LZ2020
```

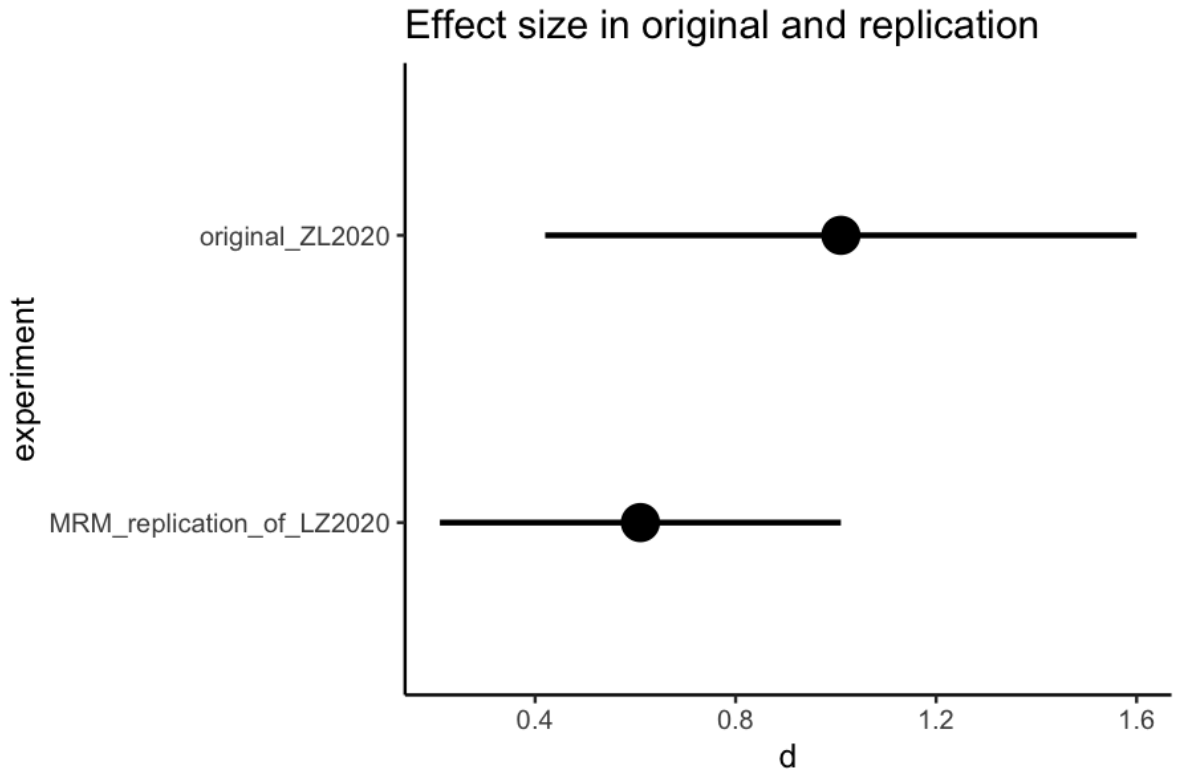
```
tidy_es <- both_es %>%
  select(experiment, d, l.d, u.d) %>%
  rename(ci_lower = l.d,
         ci_upper = u.d)

tidy_es
```

```
##                  experiment    d ci_lower ci_upper
## 1           original_ZL2020 1.01     0.42     1.60
## 2 MRM_replication_of_LZ2020 0.61     0.21     1.01
```

```
ggplot(tidy_es, aes(y = d, x= experiment)) +
  geom_pointrange(aes(ymin = ci_lower, ymax = ci_upper)) +
  ggtitle("Effect size in original and replication")
```

```
ggplot(tidy_es, aes(y = d, x = experiment)) +
  geom_pointrange(aes(ymin = ci_lower, ymax = ci_upper)) +
   ggtitle("Effect size in original and replication") +
  coord_flip()
```



Next time

- Use the statistical tools we've been talking about to discuss replication failures
- Reading:

Why Most Published Research Findings Are False

John P. A. Ioannidis

REPLICATE = Get same result with a new dataset