Statistical Foundations: Null Hypothesis Testing

24 February 2020

Modern Research Methods

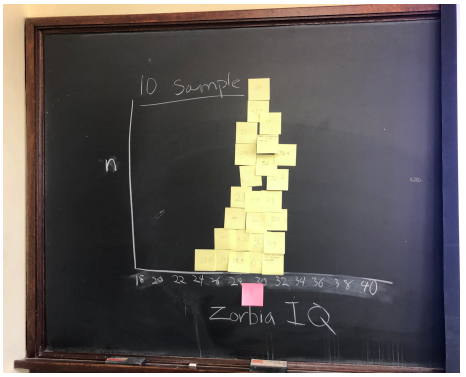
Midterm

- Handed out on Friday (2/28) at 5pm, due the following Friday (3/6) at 5pm
- Must complete on your own (using resources on the internet is fine)
- 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 this week
- Lab this Friday will be review – I won't prepare anything, you can ask me questions about any of the material we've covered so far.

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 **same value**.

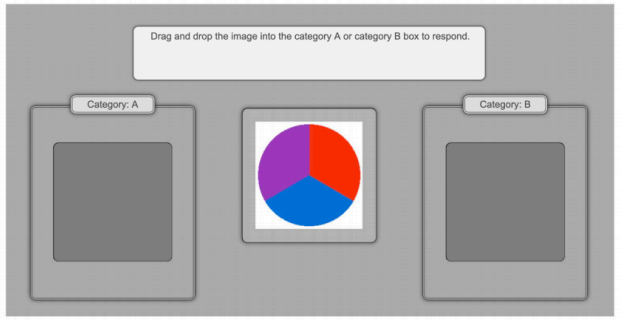
Sampling Distribution



one experiment .84

Replicating Zettersten and Lupyan (2020)

Original

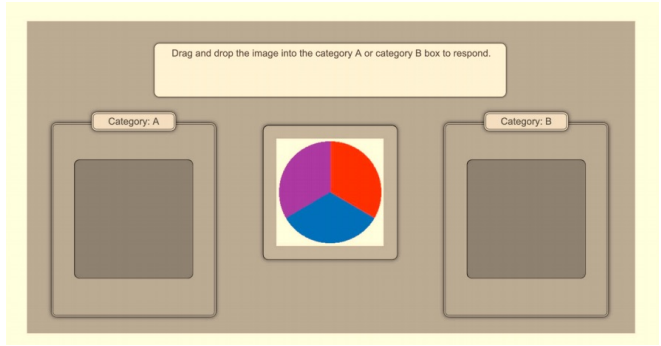


CLAIM: It's easier to learn a category when the colors are nameable.

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]



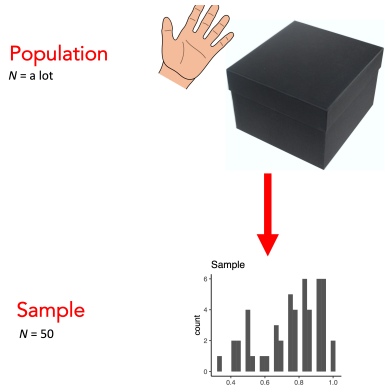
High Nameability Condition = 75%
Low Nameability Condition = 69%

Did we replicate it?

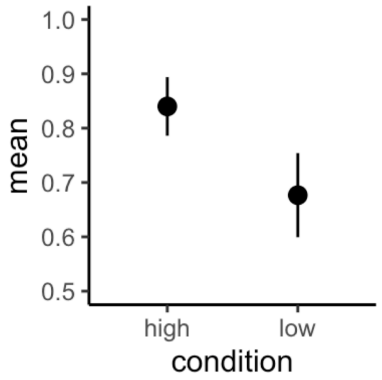
We need statistics to answer this question.

Last week:

Sampling



Confidence Intervals (CI)



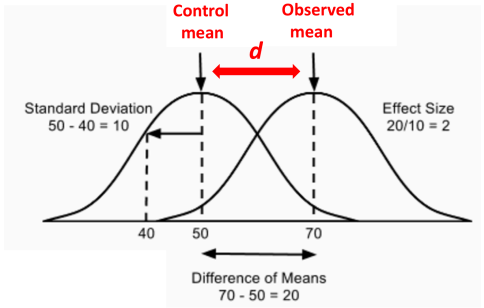
This week:

Null Hypothesis Testing

Are the means different? (yes/no)

Effect Sizes

How different are the means?



Review: Calculating and plotting CIs

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")
```

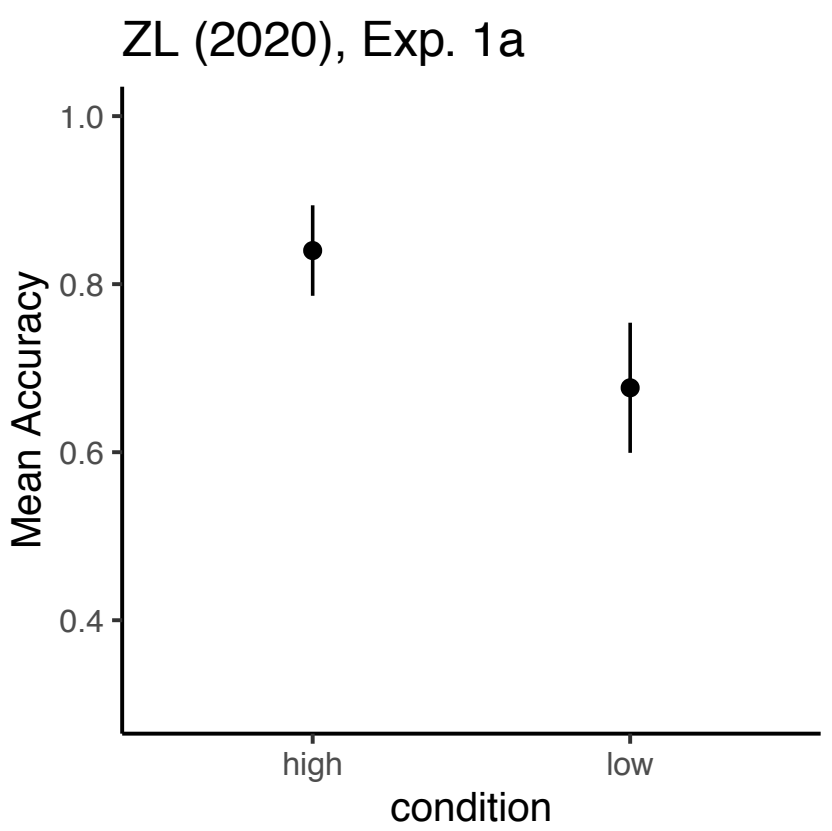
Review: Calculating and plotting CIs

```
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)
```

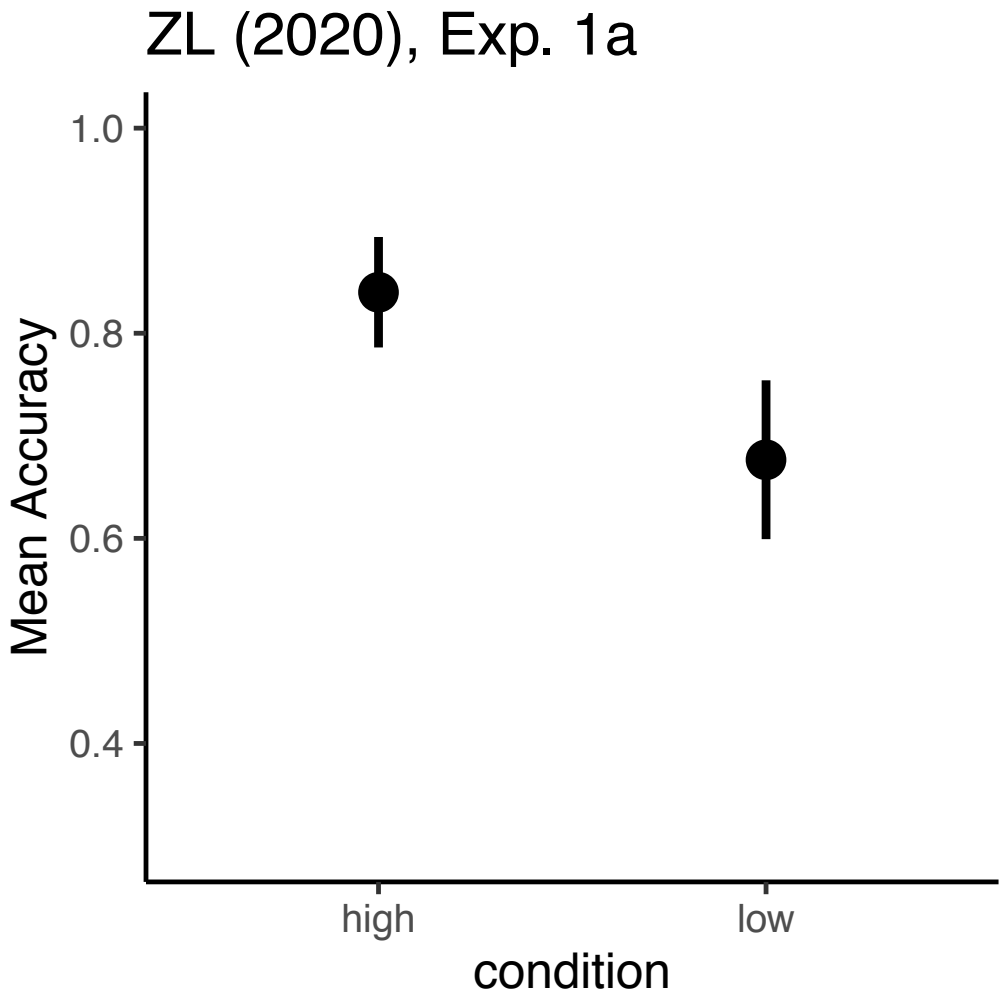
Review: Calculating and plotting CIs

```
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)
```



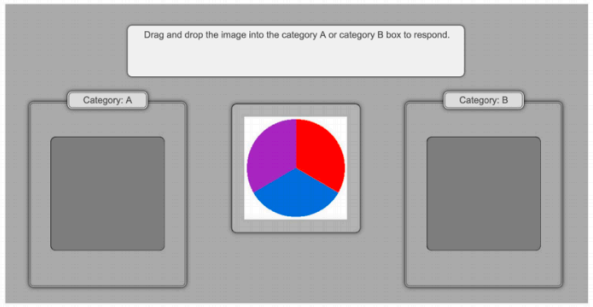
Review: Interpreting CIs

- What do CIs tell us?
- Range of plausible values for each condition.
- What is meant by *plausible*?
- What is an example of a plausible value for the high/low condition?
- What is an example of an implausible value for the high/low condition?



Today: Are the two means different from each other? (yes/no)

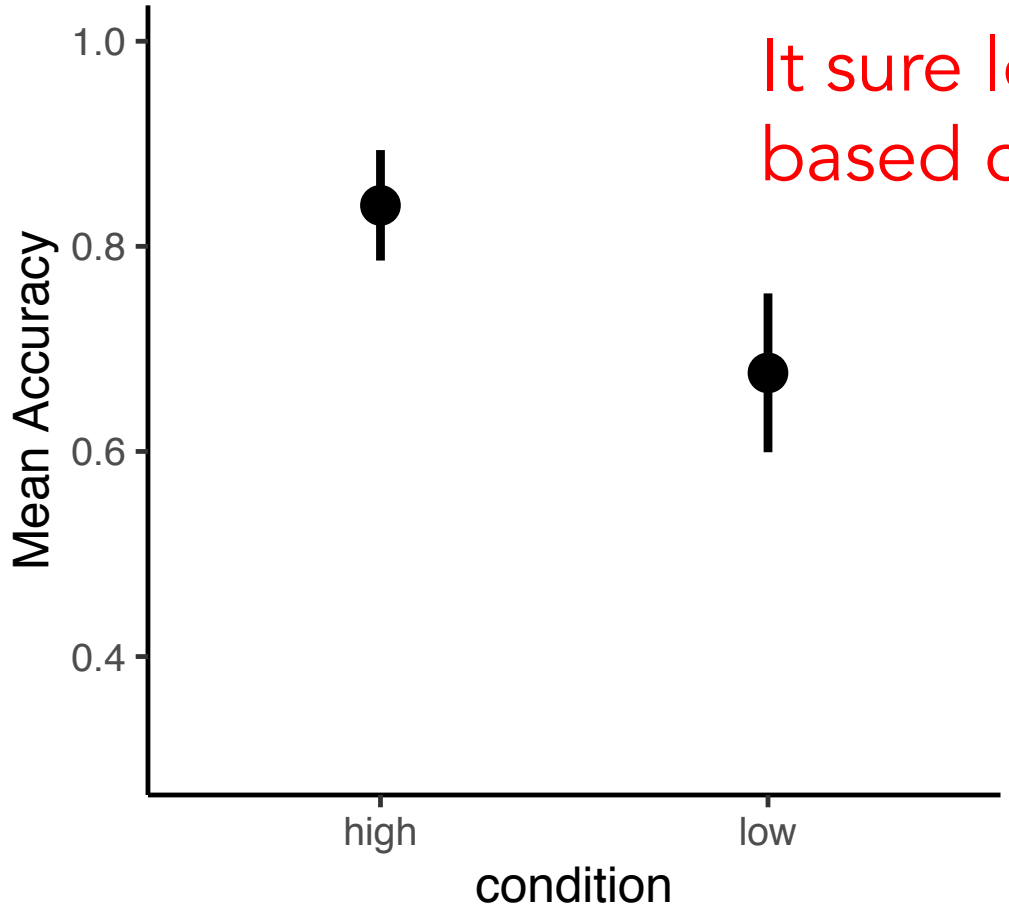
Original



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

ZL (2020), Exp. 1a



It sure looks like it based on the CIs!

Hypothesis Testing

- We're going to formally test the hypothesis that the mean in the high condition is higher than in the low condition.
- This is a **statistical hypothesis** (vs. research hypothesis)

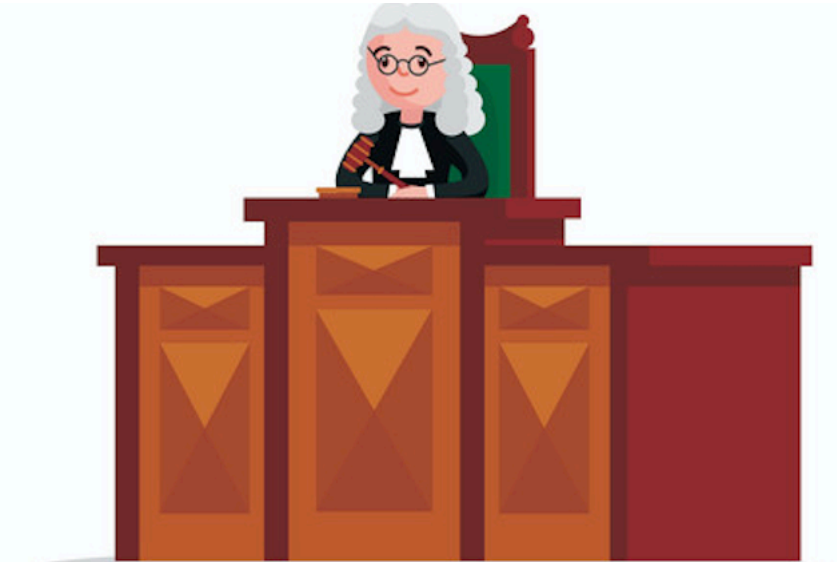
- Statistical Hypothesis: Mean accuracy in the high nameability condition is greater than the mean of accuracy in the low nameability condition.
- Research Hypothesis: It's easier to learn a category when the colors are nameable, or, "nameability helps category learning."

Testing the statistical hypothesis

- By focusing on the **null hypothesis**
- Null hypothesis: "nothing interesting going on"

Null Hypothesis
(Defendant)

Researcher
(Prosecutor)



Statistical Test
(Researcher)

Defendant presumed to be innocent; the researcher's job is to prove beyond a reasonable doubt that he is not innocent

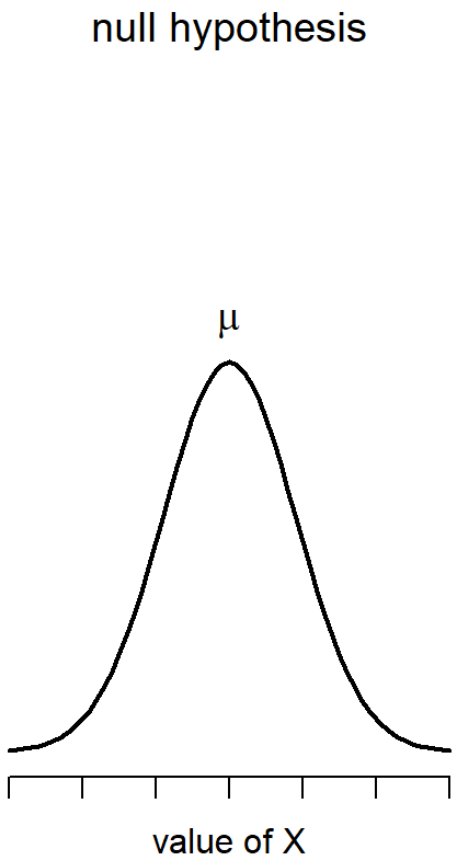
What's the null hypothesis in our case?

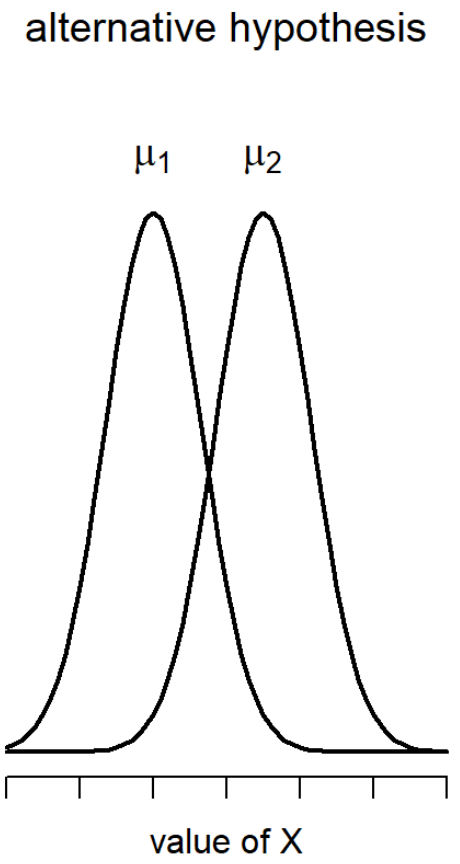
Null Hypothesis (H_0): Mean accuracy in the high nameability condition is **the same as** the mean of accuracy in the low nameability condition (i.e. they come from the same population).

Alternative Hypothesis (H_1): Mean accuracy in the high nameability condition is **greater than** the mean of accuracy in the low nameability condition. (i.e. they come from different populations)

Specifying our hypotheses about the populations formally

$$H_0: \quad \mu_1 = \mu_2$$
$$H_1: \quad \mu_1 \neq \mu_2$$





Adapted from https://learningstatisticswithr.com/, by Danielle Navarro

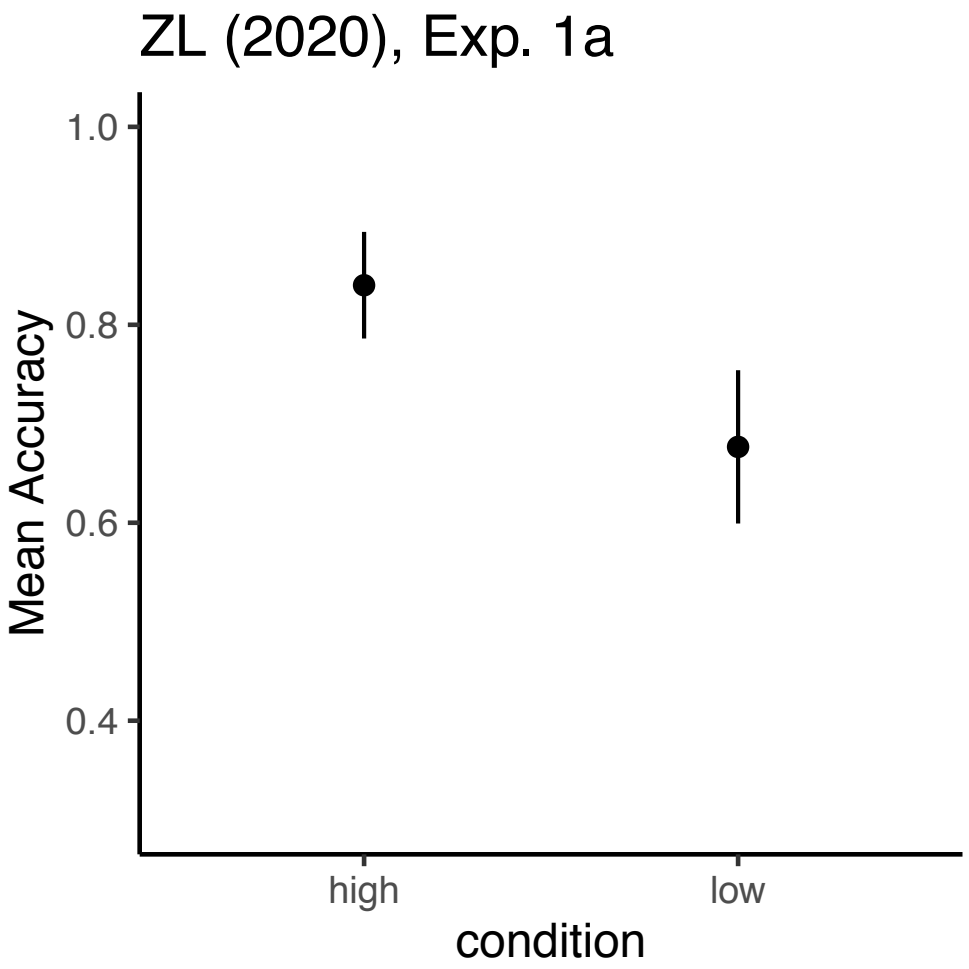
Two types of errors we can make

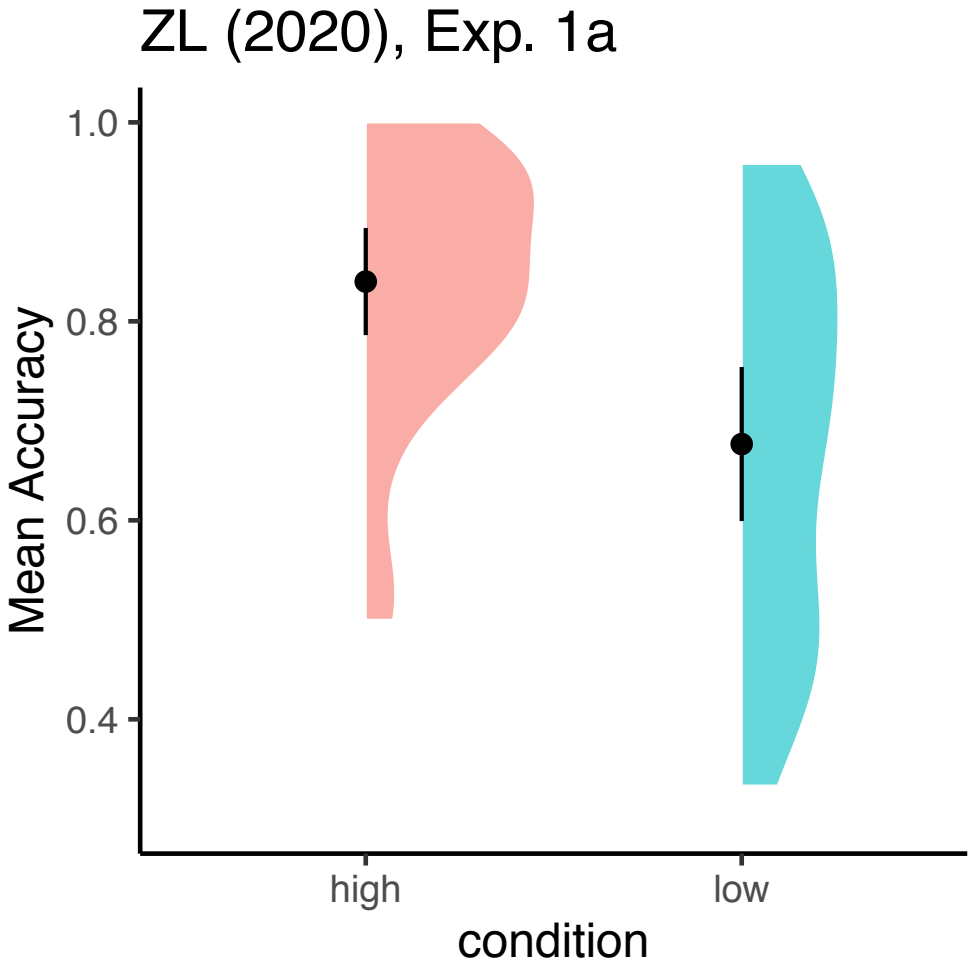
Statistical tests designed to keep rates of Type I error low

	Acquit	Convict
Innocent	correct	Error!
Guilty	Error!	correct

	Retain H_0	Reject H_0
H_0 is true	correct	Error (Type I)
H_0 is false	Error (Type II)	correct

Do these means come from the same or different population distributions?

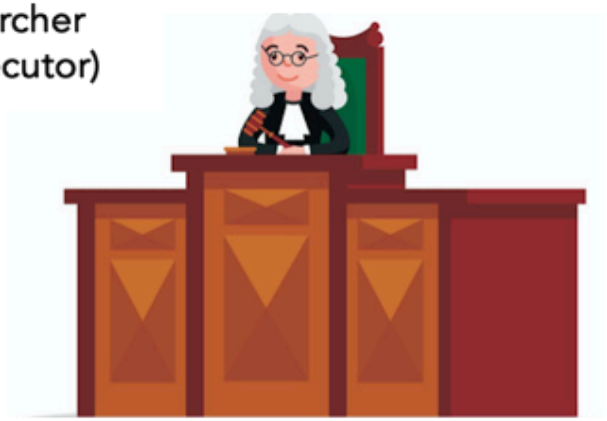




Do these means come from the same or different population distributions?

Null Hypothesis
(Defendant)

Researcher
(Prosecutor)



Statistical Test
(Researcher)

- **Two sample *t*-test** gives us a formal way to answer this question.
- Assigns **probability** to our data if the null hypothesis were true (***p*-value**)
- High p-value -> High likelihood of observing our data under the null hypothesis (i.e. means come from same population distribution)
- Low p-value -> Low likelihood of observing our data under the null hypothesis (i.e. means come from different populations)

p-values

- How improbable does the data have to be before we reject the null hypothesis?
- Decide on a p-value threshold, alpha
- In psychology, alpha is typically .05 (only 5% chance that the null hypothesis is true)
- But, alpha could be any value…

Reasoning about *p*-values

p-value = probability of our data if null hypothesis is true

alpha = threshold for rejecting null hypothesis

Talk to the person next to try to answer the following two questions:

1. **What happens to the Type I and Type II error rate if alpha is bigger than .05?**
2. **What happens to the Type I and Type II error rate if alpha is smaller than .05?**

	Acquit	**Convict**
Innocent	correct	Error!
Guilty	Error!	correct

	Retain H_0	**Reject H_0**
H_0 is true	correct	Error (Type I)
H_0 is false	Error (Type II)	correct

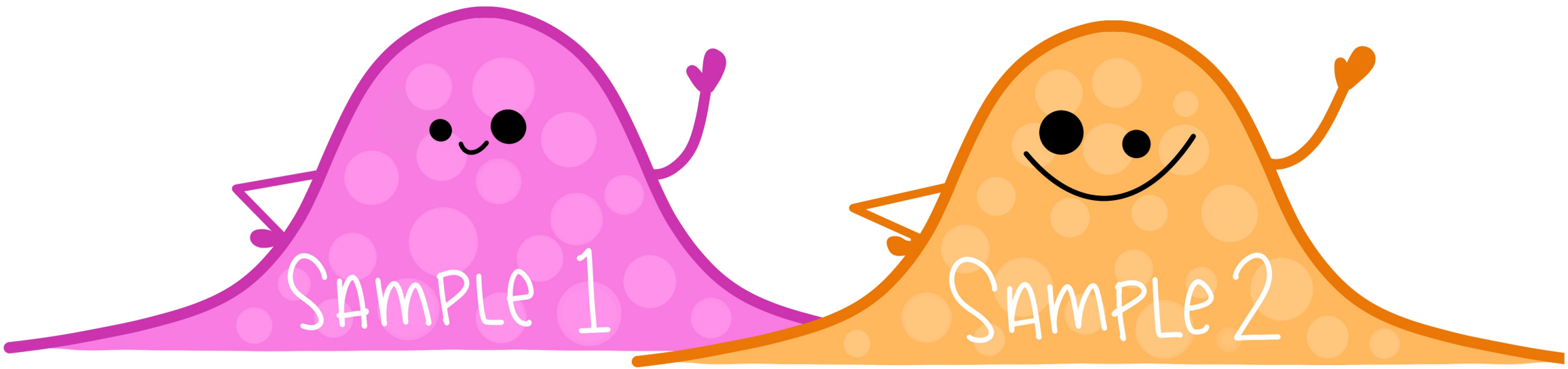
Calculating a *p*-value

- The way you calculate a *p*-value with a *t*-test is deeply related to how we calculated confidence intervals last week.
- But, the details are beyond the scope of this course.
- If you're curious, see readings or take an intro stats class.
- What's important for this course is that you have an intuition for hypothesis testing and what *p*-values are.

2-SAMPLE T-TESTS

@allison_horst

teaching assistants:

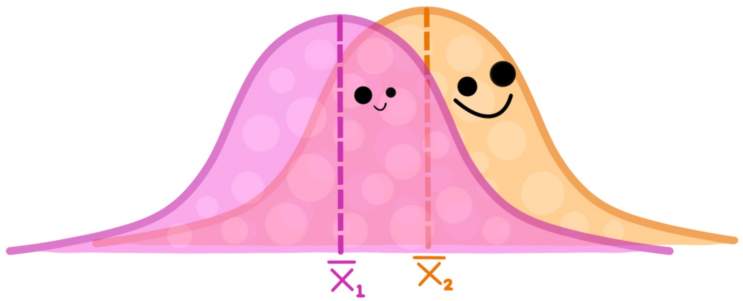


Artwork by @allison_horst

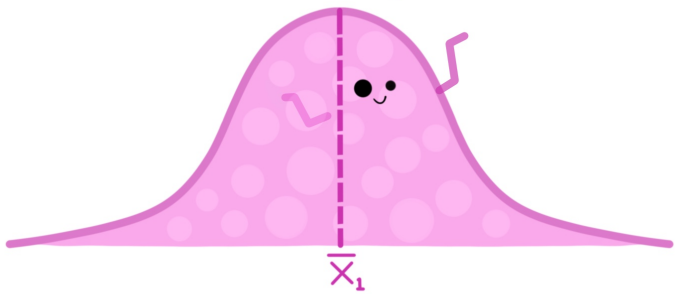
LET'S START HERE: if random samples are drawn from populations w/ the same mean...

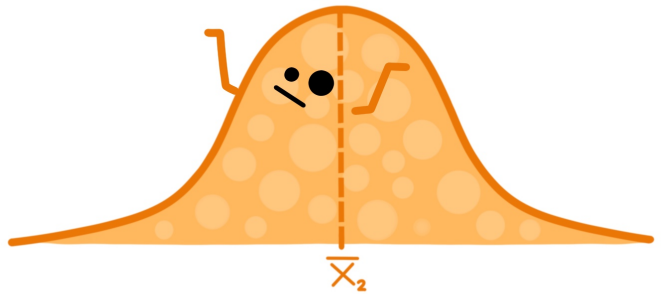
(i.e. the same population)

Then it is more likely that the 2 sample means will be close together...



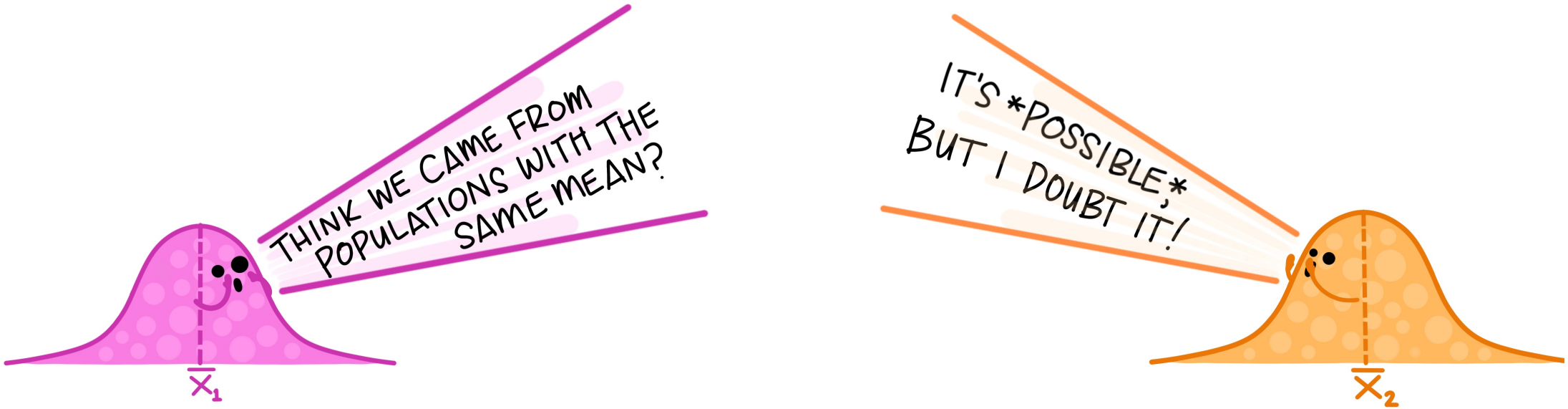
...and it is less likely (but always possible!) that the sample means will be far apart.





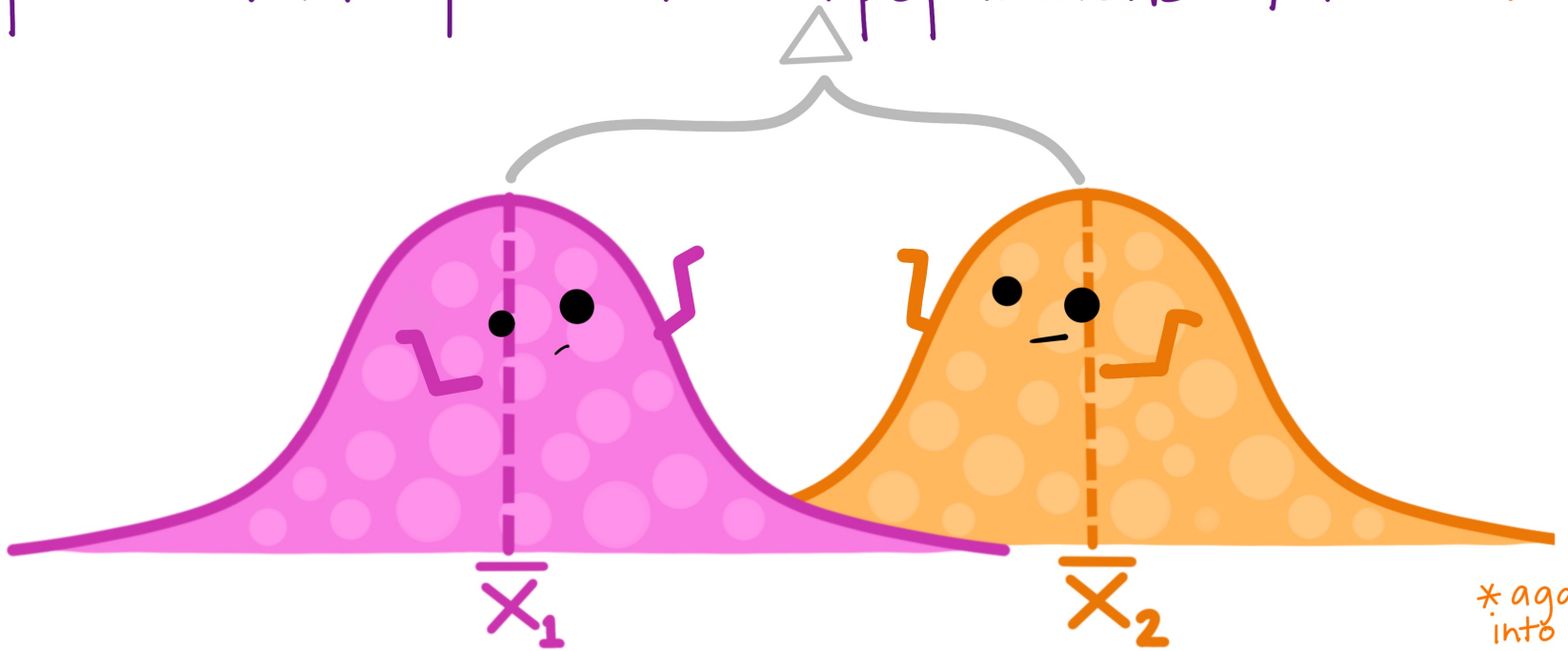
in OTHER WORDS... The more different the sample means are,* the less likely it is they were drawn from populations w/ the same mean.

*(when taking into account sample spread + size, & assuming we've randomly sampled



So for our 2 random samples, we ask:

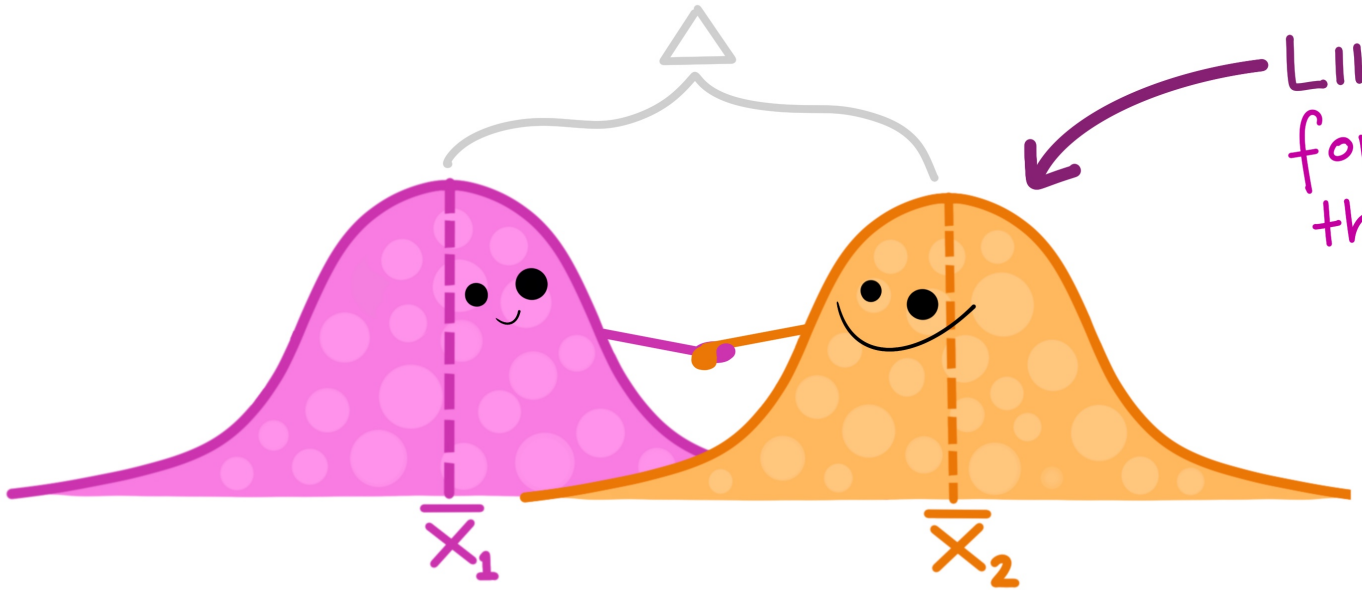
WHAT IS THE PROBABILITY OF GETTING 2 SAMPLE MEANS THAT ARE AT LEAST THIS DIFFERENT,* if they were actually drawn from populations w/ the same mean?



*again, when taking into account sample spread & size, + assumptions...

That's our p-value!

WHAT IS THE PROBABILITY OF GETTING 2 SAMPLE MEANS THAT ARE AT LEAST THIS DIFFERENT, if they were actually drawn from populations w/ the same mean?



LIKE: If a 2-sample t-test for these samples yields p=0.03, that means there is a 3% chance of getting means that are at least this different, if they're drawn from populations with the same mean.

Question:

WHEN DO WE HAVE ENOUGH EVIDENCE TO SAY THERE IS A SIGNIFICANT DIFFERENCE?

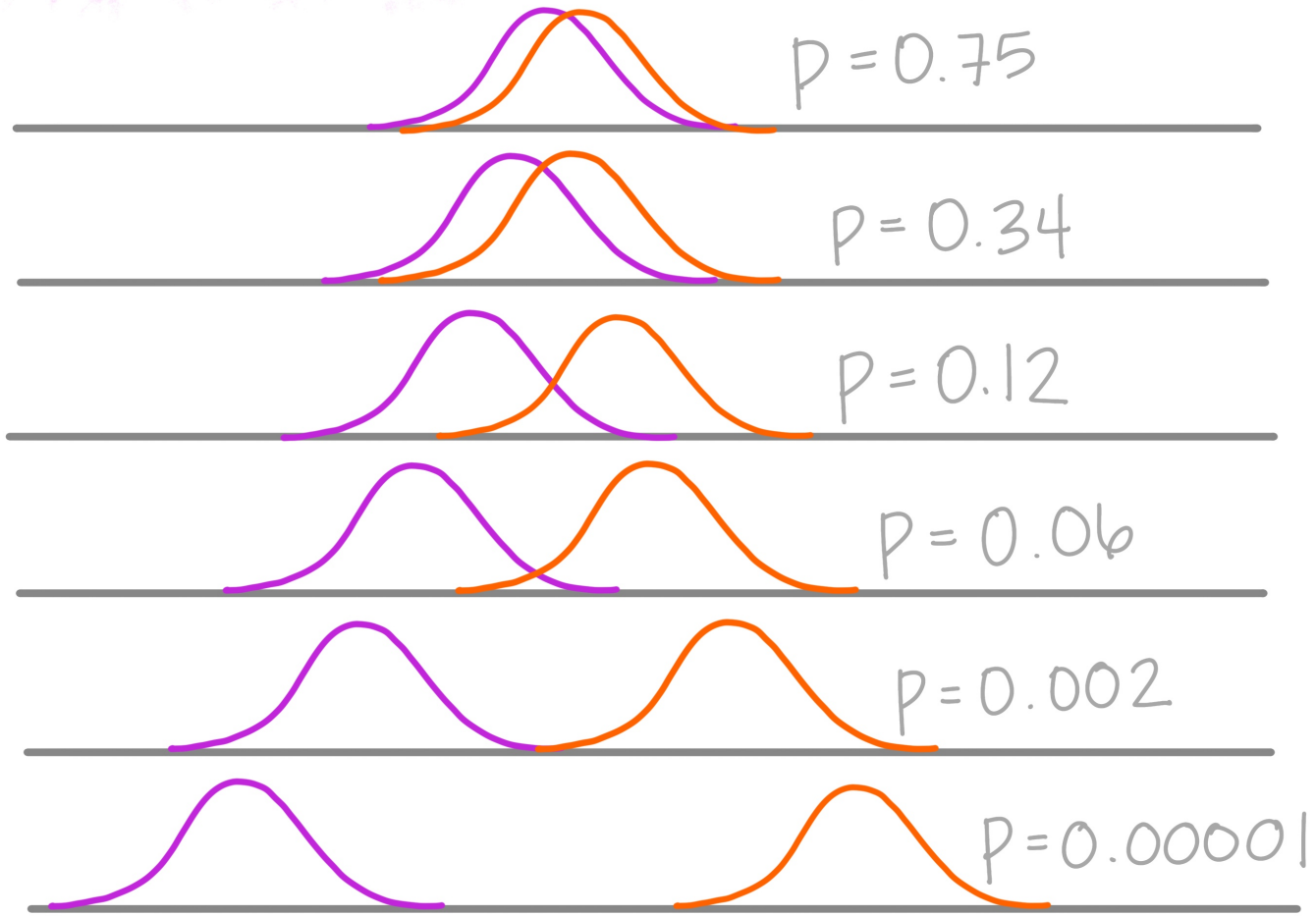
Answer:

WHEN OUR P-VALUE IS BELOW OUR SELECTED SIGNIFICANCE LEVEL (α), USUALLY (BUT NOT ALWAYS) = 0.05.

Which means:

IF THE PROBABILITY (p-value) OF FINDING AT LEAST OUR DIFFERENCE IN SAMPLE MEANS (IF THEY WERE DRAWN FROM POPULATIONS WITH THE SAME MEANS) IS LESS THAN 5%, THAT'S ENOUGH EVIDENCE FOR US TO DECIDE THEY ARE LIKELY FROM POPULATIONS WITH UNEQUAL MEANS.

P-VALUES, SCHEMATICALLY:



Higher p-values

(HIGHER PROBABILITY OF 2 SAMPLE MEANS BEING AT LEAST THIS DIFFERENT, IF DRAWN FROM POPULATIONS WITH THE SAME MEAN) = LESS EVIDENCE OF DIFFERENCES BETWEEN POPULATION MEANS

Lower p-values

(LOWER PROBABILITY OF 2 SAMPLE MEANS BEING AT LEAST THIS DIFFERENT, IF DRAWN FROM POPULATIONS WITH THE SAME MEAN) = MORE EVIDENCE OF DIFFERENCES BETWEEN POPULATION MEANS

Calculating a *p*-value in R

```
low_values <- ms_by_overall %>%
  filter(condition == "low") %>%
  pull(prop_right) # "pulls" the values from a column out of the data frame

low_values
```

```
##  [1] 0.8750000 0.7083333 0.9583333 0.6666667 0.4583333 0.9166667 0.7500000
##  [8] 0.8750000 0.7500000 0.9166667 0.8750000 0.5833333 0.4583333 0.4166667
## [15] 0.3333333 0.4166667 0.5000000 0.6666667 0.7916667 0.5000000 0.7916667
## [22] 0.8750000 0.6250000 0.5000000 0.7083333
```

```
high_values <- ms_by_overall %>%
  filter(condition == "high") %>%
  pull(prop_right)

high_values
```

```
##  [1] 0.9583333 1.0000000 0.8333333 0.8333333 0.7500000 0.9166667 0.6666667
##  [8] 0.7916667 0.9166667 0.8333333 0.7916667 0.8333333 0.9166667 0.7500000
## [15] 0.9583333 0.8333333 0.9583333 0.5416667 0.9583333 0.8333333 0.9583333
## [22] 0.5000000 0.7500000 0.9166667 1.0000000
```

Calculating a *p*-value in R

```
t.test(low_values, high_values)
```

```
##
##     Welch Two Sample t-test
##
## data:  low_values and high_values
## t = -3.5737, df = 42.816, p-value = 0.0008869
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.25551643 -0.07115024
## sample estimates:
## mean of x mean of y
## 0.6766667 0.8400000
```

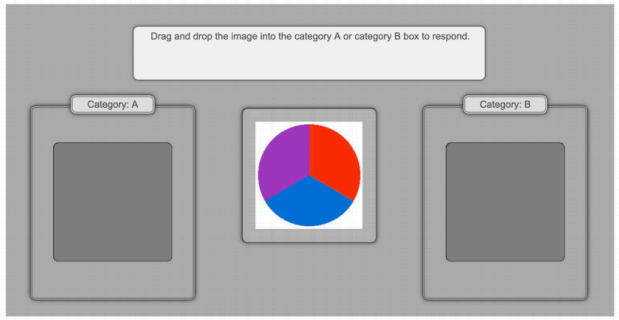
There is a .0008 chance that the null hypothesis is true.

.0008 < .05

REJECT null hypothesis

Replicating Zettersten and Lupyan (2020)

Original

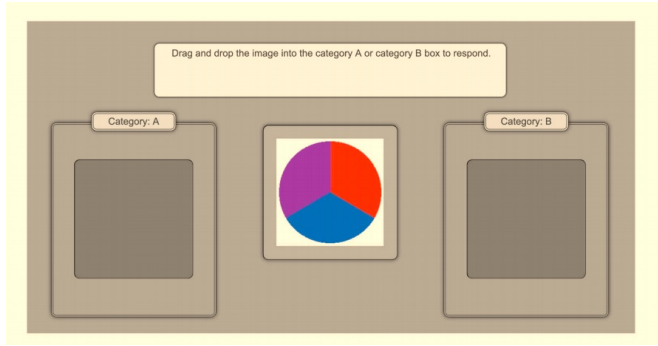


CLAIM: It's easier to learn a category when the colors are nameable.

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?

Our "replication" data

Here are the data from our "replication" in which we ran 50 participant in each of the two conditions (I made the data up).

```
ms_by_overall_replication <- read_csv("mrm_replication_data.csv")
```

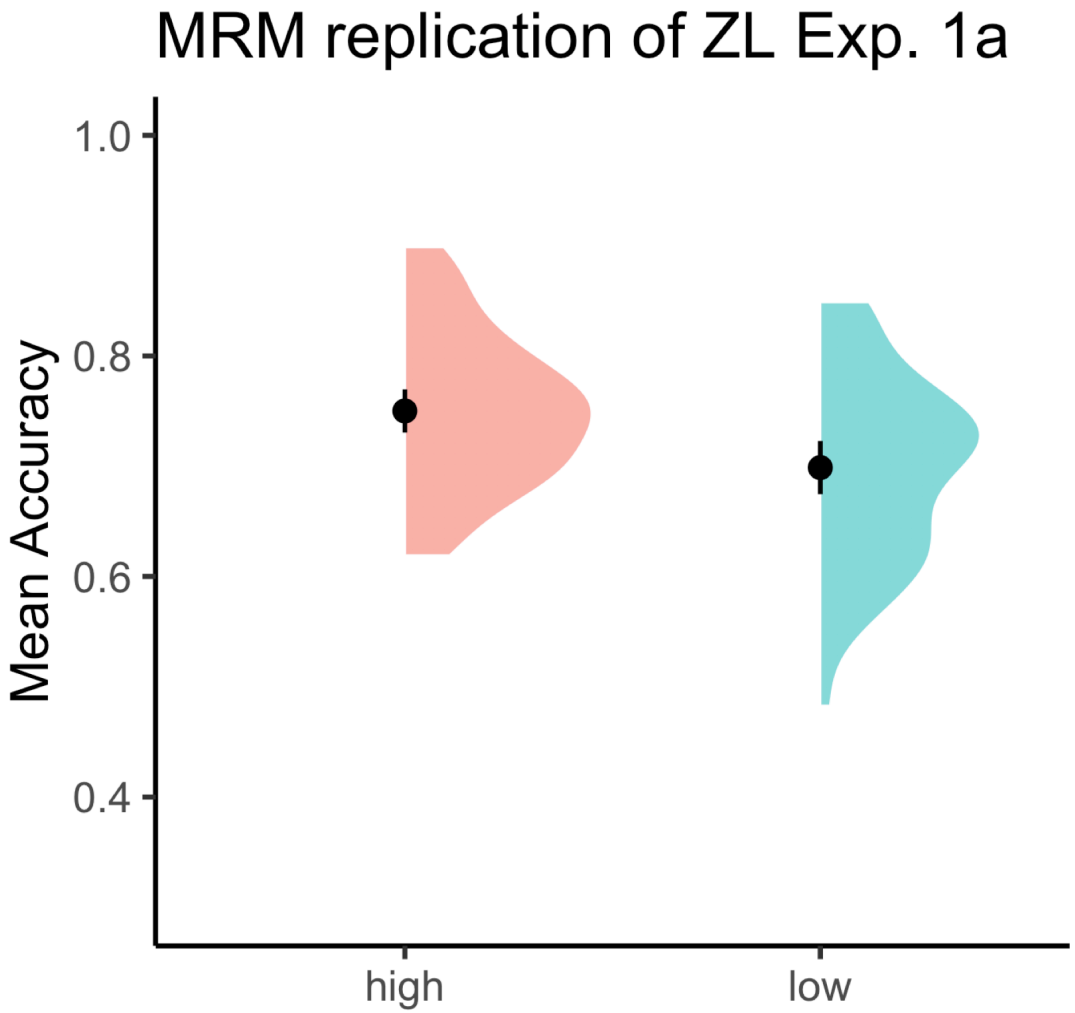
experiment	subject	condition	prop_right
MRM_replication_of_LZ2020	1	high	0.7985325
MRM_replication_of_LZ2020	2	high	0.8861798
MRM_replication_of_LZ2020	3	high	0.7950746
MRM_replication_of_LZ2020	4	high	0.7331276
MRM_replication_of_LZ2020	5	high	0.8410513

Let's calculate and plot CIs for our replication

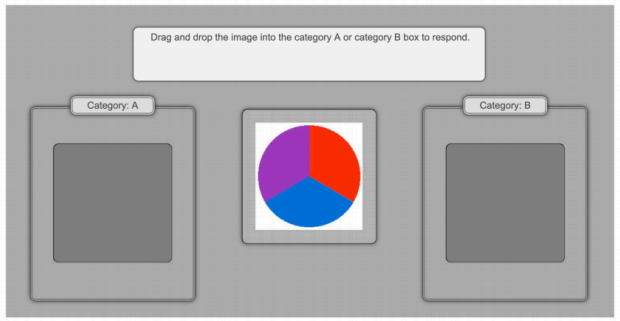
```
means_by_condition_replication <- ms_by_overall_replication %>%
  group_by(condition) %>%
  summarize(mean = mean(prop_right),
            sd = sd(prop_right),
            n = n()) %>%
  mutate(ci_range_95 =  qt(1 - (0.05 / 2), n - 1) * (sd/sqrt(n)),
         ci_lower = mean - ci_range_95,
     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

means_by_condition_replication
```

```
## # A tibble: 2 x 7
##   condition  mean     sd     n ci_range_95 ci_lower ci_upper
##   <chr>     <dbl>  <dbl> <int>       <dbl>    <dbl>    <dbl>
## 1 high      0.750 0.0692    50      0.0197    0.730    0.770
## 2 low       0.699 0.0848    50      0.0241    0.675    0.723
```

Our "replication" results

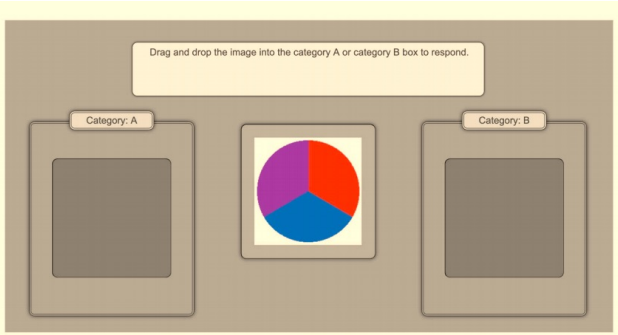


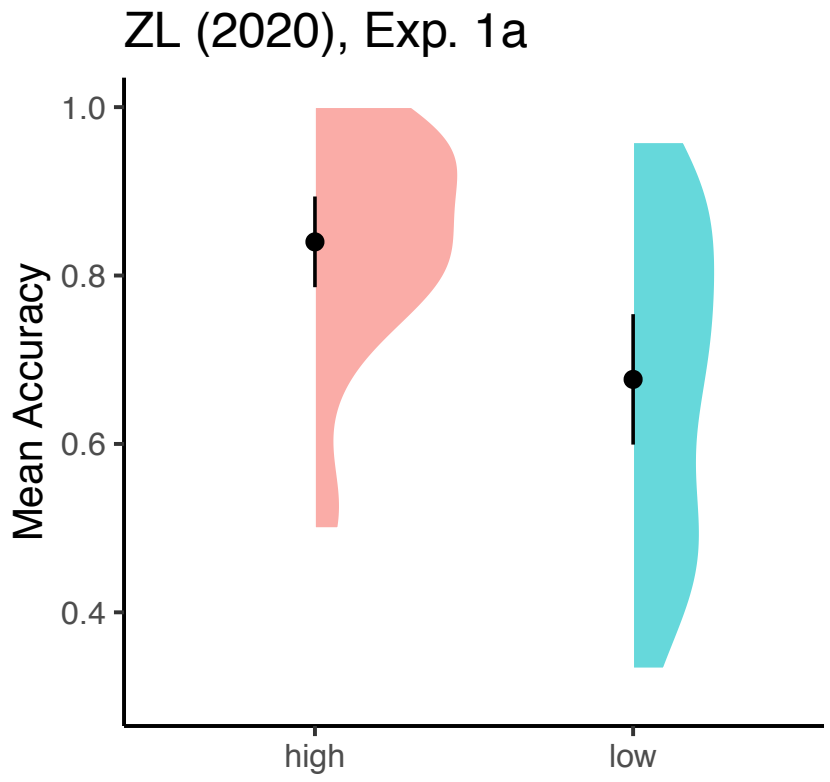
Original

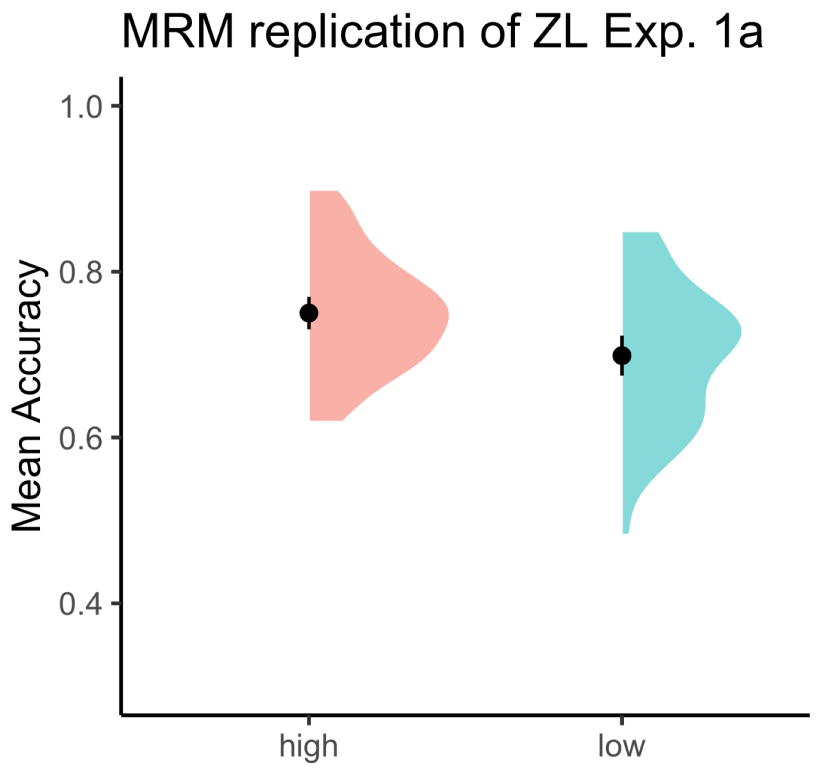


Replication

[Us]







Let's calculate a *p*-value for our replication

```
low_values_replication <- ms_by_overall_replication %>%
  filter(condition == "low") %>%
  pull(prop_right) # "pulls" the values from a column out of the data frame

high_values_replcation <- ms_by_overall_replication %>%
  filter(condition == "high") %>%
  pull(prop_right)
```

```
t.test(low_values_replication, high_values_replcation)
```

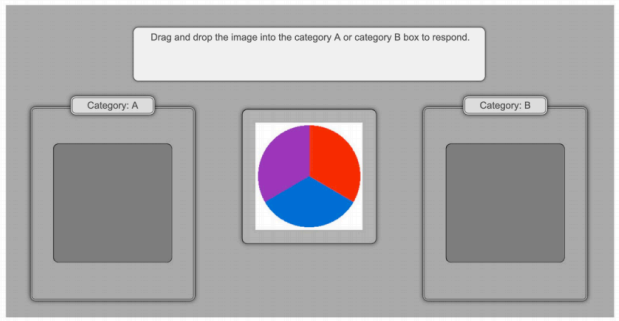
```
##
##     Welch Two Sample t-test
##
## data:  low_values_replication and high_values_replcation
## t = -3.3168, df = 94.208, p-value = 0.001294
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08210669 -0.02061654
## sample estimates:
## mean of x mean of y
## 0.6987492 0.7501109
```

There is a .001 chance that the null hypothesis is true.

.001 < .05

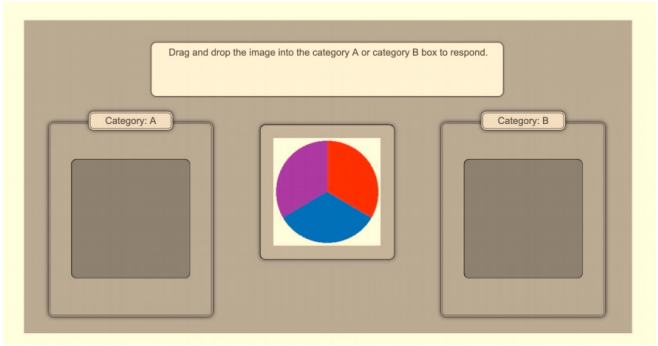
REJECT null hypothesis

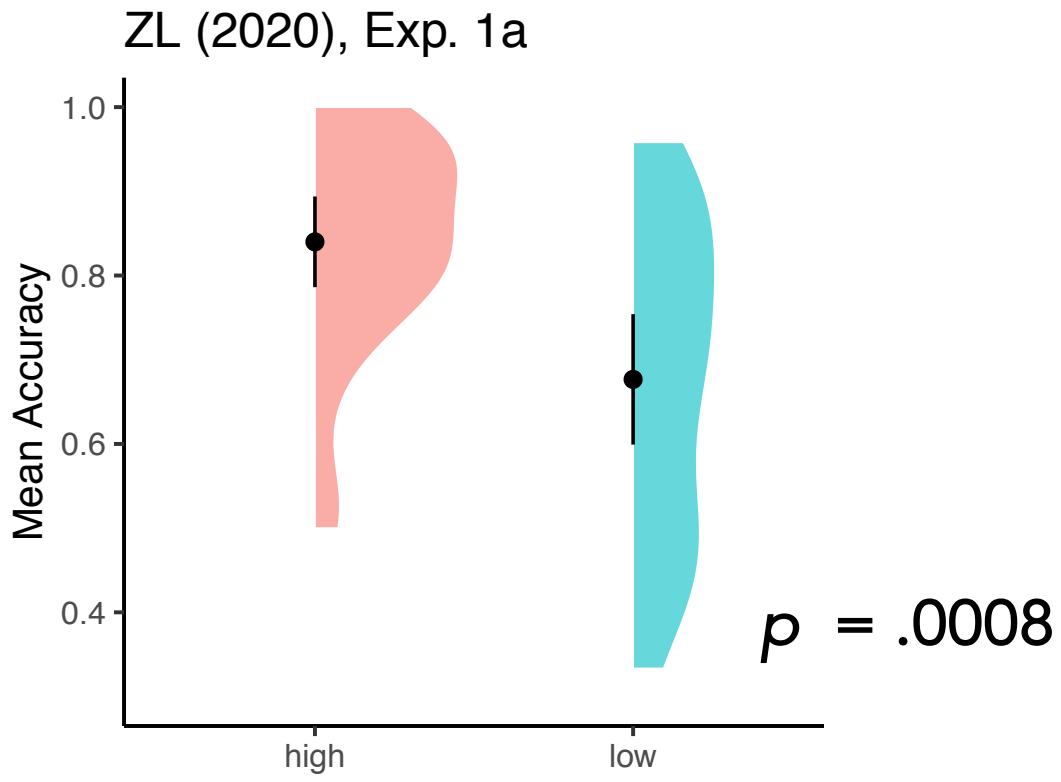
Original

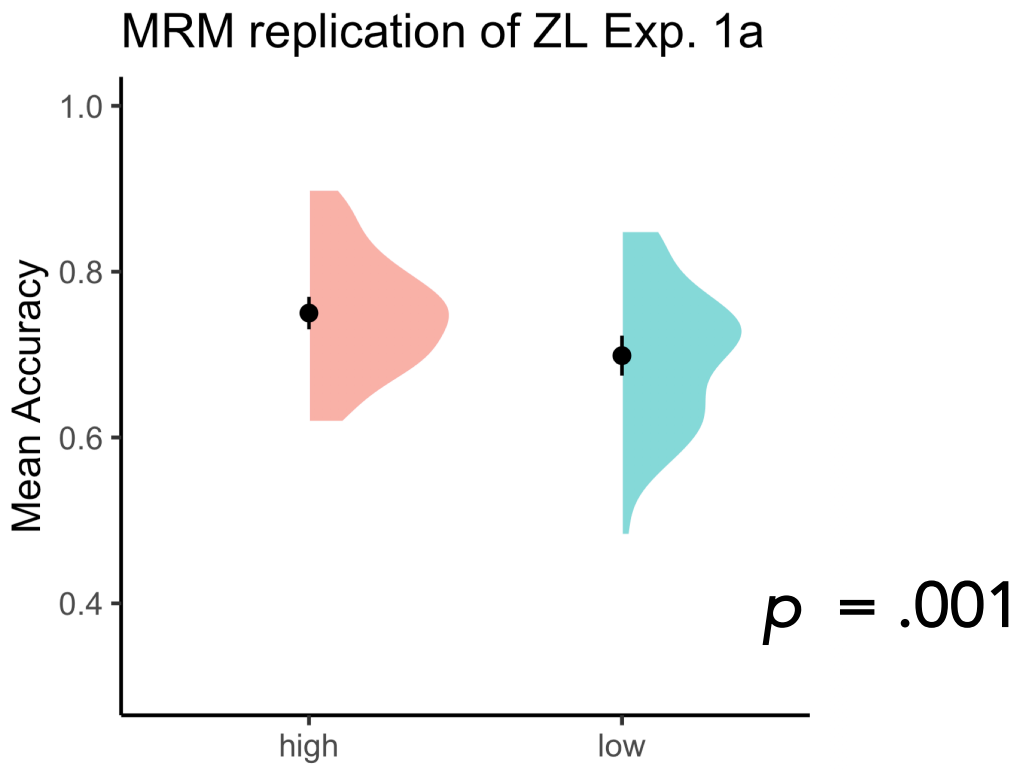


Replication

[Us]

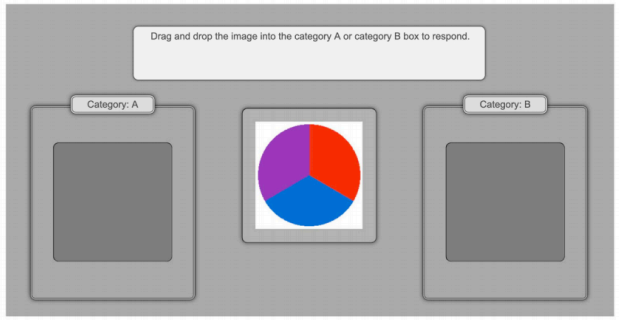






Replicating Zettersten and Lupyan (2020)

Original

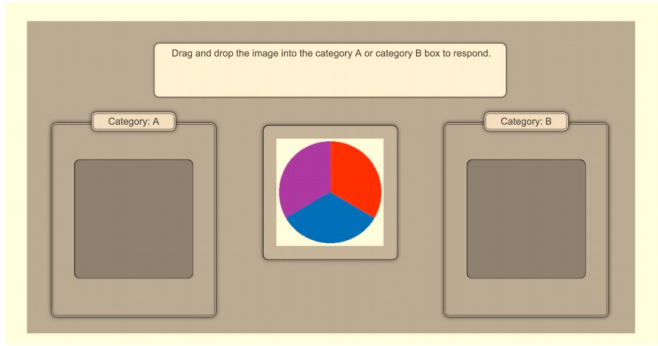


CLAIM: It's easier to learn a category when the colors are nameable.

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!

Next Time: Effect Sizes

- *p*-values tell us whether or not two means are different from each other
- Can tell us whether or not we replicated a previous experiment (yes/no)
- But, what if we also wanted to know how similar the two replications are?
- Effect sizes give us a way to quantify that.