HRI Methods

Prof. Wing Yue Geoffrey Louie and Suraj Goyal

Why?

Now that we have a system and metrics for measuring it what do we do?

This lecture will focus on common ways to conduct studies to advance the scientific basis of our field

Research Methods Overview

Experimental Method – Deliberately producing a change in one or more causal (independent variables) and measuring the effect of the change on one more effect or dependent variables.

- All other influential variables must be kept under control.
- One of the only ways to infer cause-and-effect relationships.
- Well controlled and rigorous research method

Descriptive Methods – Theorized relationship between variables are investigated but causal variable cannot be manipulated. Often just measure the independent and dependent variables and analyze the correlation between them. Referred to as a "research study" or "research project"

- For example, a researcher may be interested on the relationship of number of years with computer experience to one's ability to a control robot.

Quasi-Experiment – Same as the experimental method except you cannot control all the influential factors

- Often reduces the validity of the study

Experimental Method

Experiments involve looking at the relationship between independent and the resulting change in dependent variables. Primary goal is to show that the independent variable is the only variable responsible for changes in the dependent variable.

Step 1 – Define the problem and hypotheses

- The first step is to define a hypothesis for the cause-and-effect relationship between an independent variable and a dependent variable

Example: “A robot which utilizes immediacy cues for the communication of a health intervention will be more effective”

Independent Variable: Immediacy cues (Immediacy is the perception of physical and psychological closeness between communicators)

Dependent Variable: Intervention efficacy

- The independent and dependent variables have only been defined abstractly

Experimental Method

Step 2 – Specify the Experimental Plan - Identify all the details of the experiments to be conducted

- Describe the experimental scenario, equipment to be used, tasks to be performed, the environment, and who will participate in the study.
  - Example: Turtlebot will be delivering an intervention focusing on improving the human interactants vegetable eating habits by providing a verbal reminder in the individual's kitchen three times (7AM, 12PM, and 6PM) in a day.
- Describe how the independent variable will be manipulated
  - Example: We will have two robot conditions
    1) one where robot utilizes immediacy cues including eye contact, forward body lean and orientation towards the user, continuous small upper body movements, and frequent gestures while talking
    2) one where robot utilizes no immediacy cues, so it leans backwards, avoids eye contact, doesn't use gestures, and moves rigidly
- Describe what is exactly meant by the dependent variable and how it will be measured
  - Example: Effectiveness will be measured according to the frequency the human has eaten in a day after interacting with the robot.
- An experimental design will then be selected

Experimental Method

Step 3 – Conduct the study

- Recruit participants, development the materials/system, and prepare to conduct the study.
- A pilot study may be conducted prior to a real study to ensure that the experiment is functioning the way it should be, measuring what it is intending, and the developed technology is reliable.
- Full study is then conducted and data is gathered

Step 4 – Analyze the data

- Inferential statistics are then utilized to analyze the data
- For the described scenario, we would have the frequency of vegetables eaten by each participant after interacting with one of the robot conditions. One set of data with no immediacy cues and one set of data with immediacy cues.
- The inferential statistics would then be utilized to see if there is a significant difference with the two conditions

Experimental Method

Step 5 – Draw conclusions

- After the statistical analysis you can draw conclusions from the cause-and-effect relationships between the variables
  - Verify if hypotheses were supported (Immediacy cues improved intervention effectiveness)
  - Provide explanations for **why** such results were attained? Go beyond the obvious.
    - Immediacy cues may be beneficial because:
      - The human does not feel as if they are being commanded so they feel as if they have a choice
      - They increase user perception of trust towards the robot

Experimental Designs

Multiple approaches to running an experiment with their respective advantages and disadvantages

Two-Group Design

- One independent variable with only two conditions/levels.
- A control group of subjects gets no treatment (e.g. No robot)
- Experimental group of subjects get “some” amount of the independent variable (e.g. Robot)
- Doesn’t make sense in our previous example

Experimental Designs

Multiple Group Designs

- Two levels/conditions may not be sufficient in some cases to evaluate a hypothesis and we may want to test many levels of an independent variable
- For example, if we want to evaluate three levels of immediacy cues (No cues, some cues, all cues) we would need three groups.
- This is important for scenarios where curvilinear relationships between independent and dependent variables are expected (e.g. if we expect that increasing the amount of immediacy cues will lead to exponential increases in broccoli eating)

Experimental Designs

Factorial Design

- A factorial design systematically examines multiple independent variables with multiple levels
- All combinations of “factors” are evaluated
- Enables the assessment of independent variables individually and the relationship between independent variables
- Benefits include allowing the variation of more variables and better captures the complexity in the real-world

	Cute Robot	Transformer Robot
High Immediacy	HI x CR	HI x TR
Low Immediacy	LI x CR	LI x TR

Experimental Designs

Between-Subjects Design

- Between-subjects design refers to when different groups of subjects are used to investigate the different levels/conditions of an independent variable
- Multi-level, two-group, and factorial designs can all be subsets of a between-subjects design
- One of the uses of a between-subjects design is when the same subject cannot participate in both conditions because it would affect the results
  - For example, if a person was conducting the same USAR task in the same scenario with the victims in the same location you would not be able to use the same condition on both subjects due to a learning effect.

Experimental Designs

Within-Subjects Design

- Within-subjects design refers to when the same subject participates in all experimental conditions
- Performance would then be compared across the participant
- The advantage of a within-subjects design is that it is easier to find statistically significant differences between conditions because there are less probability of confounding factors
- Another advantage is you often require less participants

Experimental Designs

Mixed Designs

- In factorial designs it is also possible to utilize between-subjects design for one independent variable and a within-subjects design for another independent variable

- For example:
  - One independent variable could be the level of training provided to an operator for an USAR interface. One group of subjects is not trained and another group of subjects is trained
  - A second independent variable could be the # of victims being searched which could have 3 levels (1 victim, 2 victims, 3 victims). All subjects are tasked with all 3 of these levels.

Additional Considerations

Multiple Dependent Variables – Systems are often complex so it often feasible and efficient to simultaneously measure multiple dependent variables at the same time

Selecting the Apparatus and Context – After specifying the experimental design for dependent and independent variables we need to decide the tasks and contexts which a person will be performing. The more generalizable the results the better.

Additional Considerations

Controls

- All extraneous variables have potential to interfere with the causal relationship (AKA confounds).

- For example, in a driving scenario measuring the effect of cellphone usage you wouldn't want only older adults in the cellphone condition and adults in the non-cellphone condition.

- Randomizing assignment of individuals reduces the effect of confounding variables

Additional Considerations

Controls

- Other confounding variables beyond subjects variables should also be controlled such as environment, technology (e.g. same vehicle style), and tasks (e.g. same driving course)

- For within-subject studies there is also an additional confound associated with the order in which conditions are presented to the subjects (e.g. carryover, practice, fatigue, context effect).
  - For example, if you test a participant on 5 different control modalities (e.g. mouse, keyboard, joystick, VR, touchpad) for robot teleoperation and keep the order you present them the same they may get tired by the fifth modality every time. This leads to confounds in your results.

- Counterbalancing the presentation of conditions addresses this issue.
  - Reverse Counterbalancing – Participants receive all treatments twice: first in one order and next in another order.
  - Complete Counterbalancing - Each possible sequence is used and each sequence is used the exact same amount of times
  - Randomized Partial Counterbalanced – Completely randomize the order of the treatments

Data Analysis

- Time to find out if there is a relationship between your independent and dependent variables.

- There are two types of statistics used to verify the hypotheses;

1) Descriptive statistics – Summarizes the dependent variable for the treatment conditions

2) Inferential statistics – Determines whether there are significant differences between experimental groups and that the relationships are not just random chance/fluctuations

Descriptive Statistics

Differences between conditions are usually described by their **mean** scores:

$$\overline{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

The spread of the scores are then described by the standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

Inferential Statistics

- Although the means of the groups may be different they could be different purely by chance.
- It is rare that humans would show the same exact performance even without changing independent variables
- Inferential statistics answers the question of whether the difference between groups was large enough that it was not only purely chance that varying the independent variable had an effect
- Statistical test for two-group design would be a T-test
- Statistical test for more than two groups would be an ANOVA

Drawing Conclusions

Type II errors – Concluding that the independent variable had no effect when it did

- This would mean a researcher evaluating a new robot control modality may erroneously conclude that it has no effect on operator performance.

$p = 0.05$ is commonly utilized as the rule of thumb for experimental analysis but this is often difficult because:

- Smaller number of subjects leads to lower statistical power and more likely to show “no significance”
- Variation in performance between different subjects or within the same subject can be large which again will result in a higher likelihood of “no significance”

Drawing Conclusions

- Statistical inference tests are based on the law of averages

Example: A cancer drug is tested and a p = 0.20 is observed.

What does this mean? There is a 1 in 5 chance that the drug had no effect

Should we give-up on the drug? No, because there is a high probability that no effect was observed by chance (Type II error)

What should we do? Run more experiments and analyze the results

- Did we not recruit enough participants?
- Did it help some people and not others?
- If it helped four out of twenty people we need to determine why it helped those four because we may be targeting the wrong population.

Drawing Conclusions

In human factors, where conclusions support design decisions and not just development and theories, researchers must be more aware of the tradeoffs between Type I and Type II errors and not blindly apply the traditional 0.05 cutoff

Statistical vs. Practical Significance

- Once chance has been ruled out ($p < 0.05$) we can with reasonable certainty say that there are differences between conditions. However, this isn't the complete story.
- Although there may be an effect the differences may not be large between conditions

Example: We compare two interfaces for controlling a robot for navigation: 1) a $1,000,000 high-resolution interface and 2) a $1 low-resolution interface

- The high-resolution interface group completes a navigation task 1% faster on average then the low-resolution interface
- With a large number of subjects this may have led to a statistically significant difference between the two groups and we could conclude that the high-quality interface is better
- What is wrong with this? Practically speaking it is not worth it to make this change if we consider the cost-benefit of having a 1% improvement.

T-test

Dependent *t*-test

- Compares two means based on related data.
- E.g., Data from the same people measured at different times.
- Data from ‘matched’ samples.

Independent *t*-test

- Compares two means based on independent data
- E.g., data from different groups of people

Significance testing

- Testing the significance of *t*
- Testing the significance of *b* in regression.

Robot Immediacy

Do humans that interact with a robot with more immediacy cues eat vegetables more frequently?

- 24 Participants

Manipulation

- Placed participants in an enclosed community riddled with hidden cameras.
- 12 participants were given a robot with high immediacy.
- 12 participants were given a robot with low immediacy.

Outcome

- measured how frequently a participant ate vegetables in a week.

Categorical predictors in the linear model

$$Y_i = (b_0 + b_1 X_{1i}) + \varepsilon_i$$

$$\text{Vegetable\_}Consumption_i = (b_0 + b_1 \text{Robot\_}Immediacy_i) + \varepsilon_i$$

Low Immediacy Group

The group variable = 0

b_0 = mean of baseline group (i.e. low robot immediacy)

$$\text{Vegetable\_}Consumption_i = (b_0 + b_1 \text{Robot\_}Immediacy_i)$$

$$\bar{X}_{Low\ Immediacy} = b_0 + (b_1 \times 0)$$

$$b_0 = (b_1 \times 0)$$

$$b_0 = 3.75$$

High Immediacy Group

The group variable = 1

*b*1 = Difference between means

$$\text{Vegetable\_}Consumption_i = (b_0 + b_1 \text{Robot\_}Immediacy_i)$$

$$\bar{X}_{High\ Immediacy} = b_0 + (b_1 \times 1)$$

$$\bar{X}_{High\ Immediacy} = (b_0 + b_1)$$

$$\bar{X}_{High\ Immediacy} = \bar{X}_{Low\ Immediacy} + b_1$$

$$b_1 = \bar{X}_{High\ Immediacy} - \bar{X}_{Low\ Immediacy}$$

Output from a linear model

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.750	.516		7.270	.000
	Robot Immediacy	1.250	.730	.343	1.713	.101

a. Dependent Variable: Vegetable Consumption

Rational for the *t*-test

Two samples are collected and the means calculated.

- These means might differ by either a little or a lot.

If the samples come from the same population, although it is possible for their means to differ because of sampling variation, we would expect large differences between sample means to occur very infrequently. Under the null hypothesis we expect means from two random samples to be very similar.

We compare the difference between the sample means that we collected to the difference between the sample means that we would expect to obtain (in the long run) if there were no effect. If the difference between the samples we have collected is larger than we would expect based on the standard error then one of two things has happened:

- There is no effect but sample means from our population fluctuate a lot and we happen to have collected two samples that produce very different means.
- The two samples come from different populations, which is why they have different means and this difference is, therefore, indicative of a genuine difference between the samples. In other words, the null hypothesis is unlikely.

The larger the observed difference between the sample means (relative to the standard error), the more likely it is that the two sample means differ because of the different testing conditions imposed on each sample.

Rationale to the *t*-test

$$t = \frac{\text{observed difference between sample means} - \text{expected difference between population means (if null hypothesis is true)}}{\text{estimate of the standard error of the difference between two sample means}}$$

The Dependent *t*-test

$$t = \frac{\overline{D} - \mu_D}{s_D / \sqrt{N}}$$

The Independent *t*-test

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Assumptions of the *t*-test

We are using a special case of the linear model, so the follow assumptions apply:

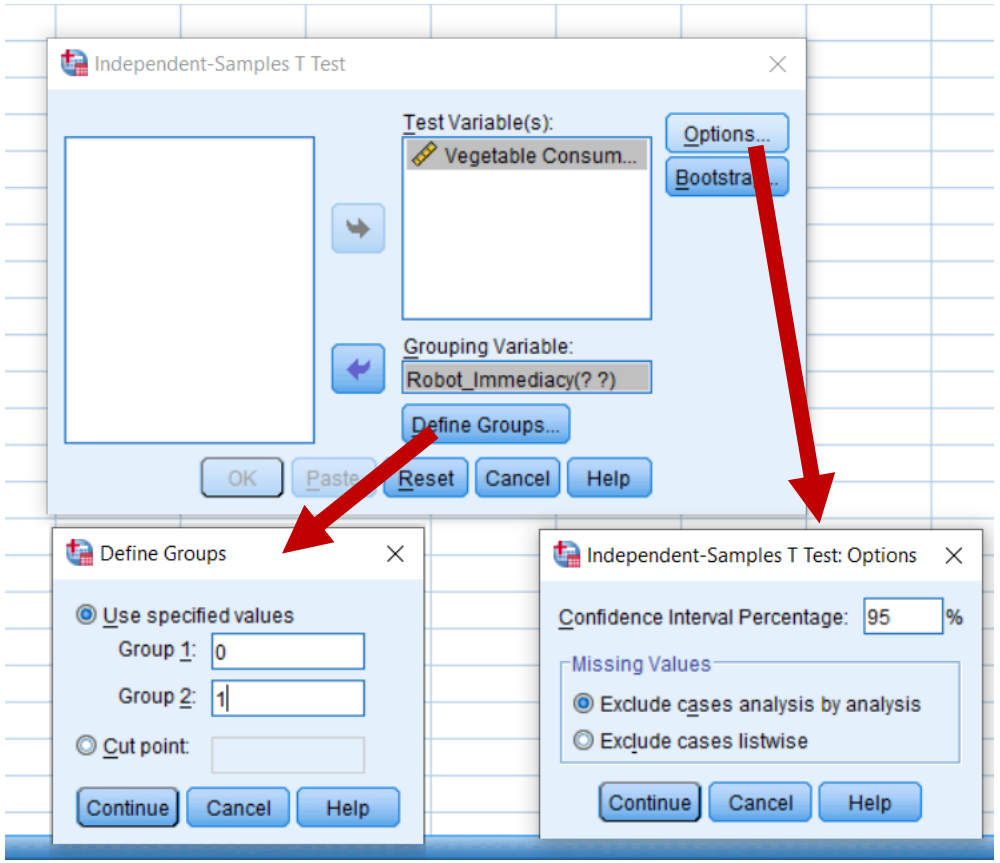
Both the independent *t*-test and the dependent *t*-test are *parametric tests* based on the normal distribution. Therefore, they assume:

- The sampling distribution is normally distributed. In the dependent *t*-test this means that the sampling distribution of the *differences* between scores should be normal, not the scores themselves.
- Data are measured at least at the interval level.

The independent *t*-test, because it is used to test different groups of people, also assumes:

- Variances in these populations are roughly equal (*homogeneity of variance*).
- Scores in different treatment conditions are independent (because they come from different people).

Independent *t*-test using SPSS Statistics



Independent *t*-test Output

Group Statistics

	Robot Immediacy		Statistic	Bootstrap[a] Bias	Bootstrap[a] Std. Error	Bootstrap[a] BCa 95% Confidence Interval Lower	Bootstrap[a] BCa 95% Confidence Interval Upper
Vegetable Consumption	Low Immediacy	N	12				
		Mean	3.75	.01	.53	2.67	4.78
		Std. Deviation	1.913	-.117	.374	1.166	2.312
		Std. Error Mean	.552				
	High Immediacy	N	12				
		Mean	5.00	-.02	.48	4.18	5.90
		Std. Deviation	1.651	-.096	.317	1.166	1.954
		Std. Error Mean	.477				

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Always use equal variances not assumed

Independent Samples Test

		Levene's Test for Equality of Variances F	Levene's Test for Equality of Variances Sig.	t-test for Equality of Means t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Vegetable Consumption	Equal variances assumed	.545	.468	-1.713	22	.101	-1.250	.730	-2.763	.263
	Equal variances not assumed			-1.713	21.541	.101	-1.250	.730	-2.765	.265

Bootstrapping output

Bootstrap for Independent Samples Test

		Mean Difference	Bootstrap[a]			
			Bias	Std. Error	BCa 95% Confidence Interval	
					Lower	Upper
Vegetable Consumption	Equal variances assumed	-1.250	.026	.720	-2.702	.070
	Equal variances not assumed	-1.250	.026	.720	-2.702	.070

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Effect sizes

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

$$r = \sqrt{\frac{-1.713^2}{-1.713^2 + 22}} = \sqrt{\frac{2.93}{24.93}} = 0.34$$

$$\hat{d} = \frac{(\bar{X}_{High\ Immediacy} - \bar{X}_{Low\ Immediacy})}{s_{Low\ Immediacy}} = \frac{5 - 3.75}{1.91} = 0.65$$

Reporting the independent *t*-test

On average, participants given a High Immediacy robot ate vegetables more frequently (M = 5, SE = 0.48), than those with a Low Immediacy robot (M = 3.75, SE = 0.55). This difference, –1.25, BCa 95% CI [–2.702, 0.070], was not significant $t(21.54) = -1.71$, $p = 0.101$; however, it did represent a medium-sized effect $d = 0.65$.

Paired-samples *t*-test

Do humans that interact with a robot with more immediacy cues eat vegetables more frequently?

- 24 Participants

Manipulation

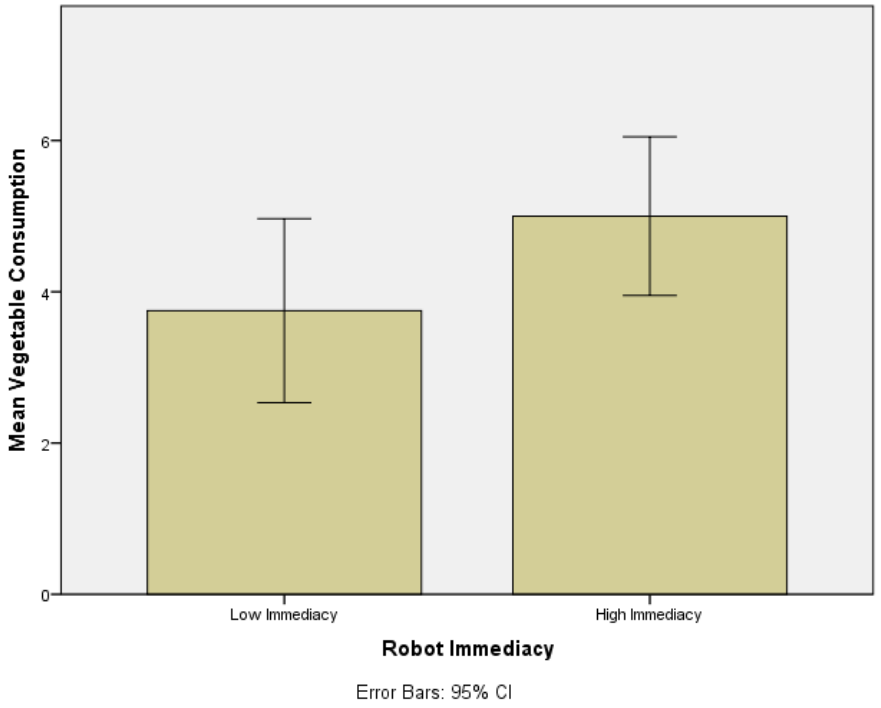
- Placed participants in an enclosed community riddled with hidden cameras.
- For first week participant, participants were given a low immediacy robot. For the second week, participants were given a high immediacy robot.

Outcome

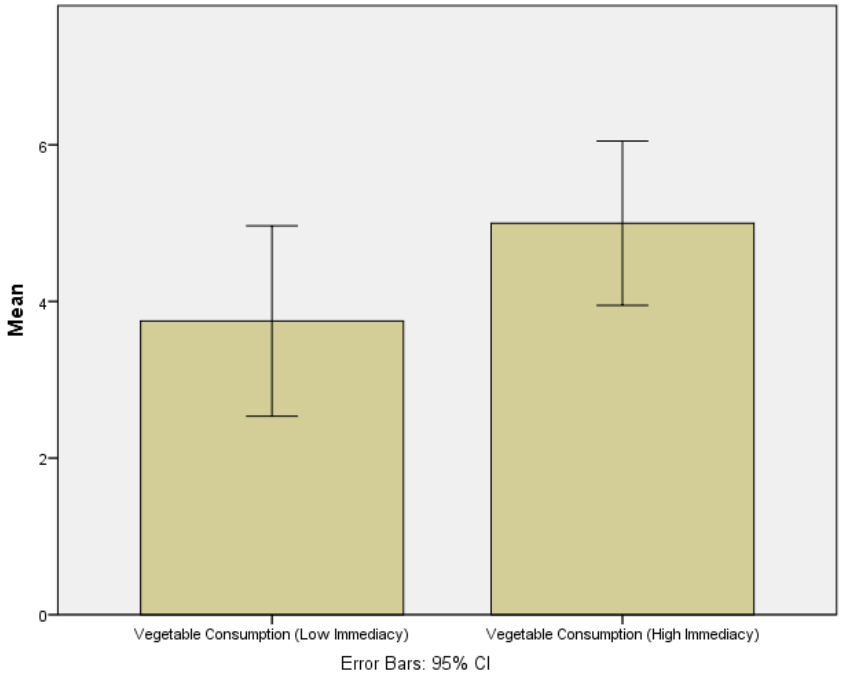
- measured how frequently a participant ate vegetables in week 1 and week 2.

Error Bar Chart for Different Studies

Confidence Intervals Generated from between-subjects study



Confidence Intervals Generated from within-subjects study

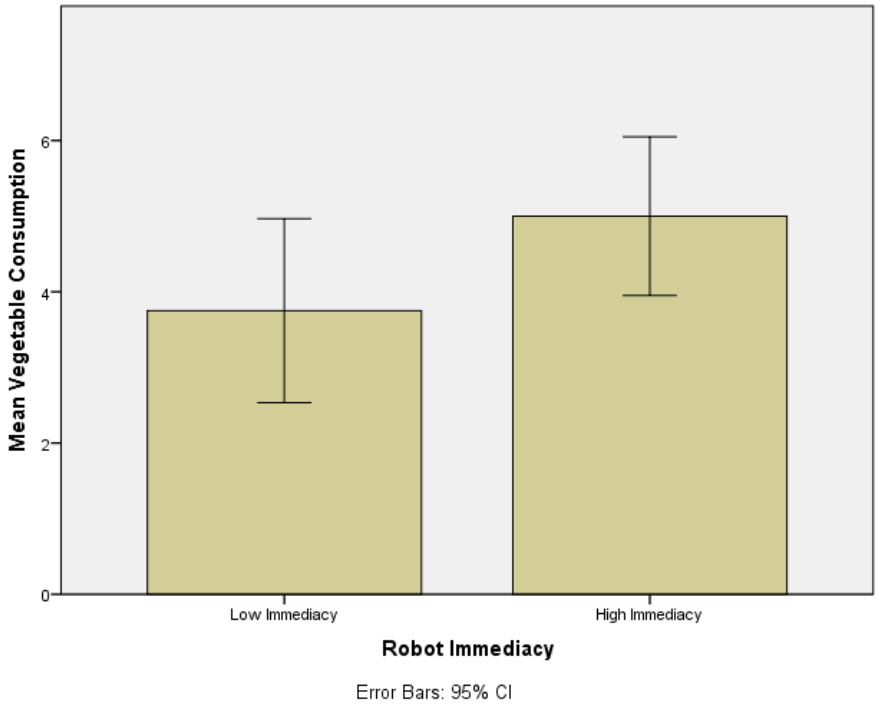


What's the difference?

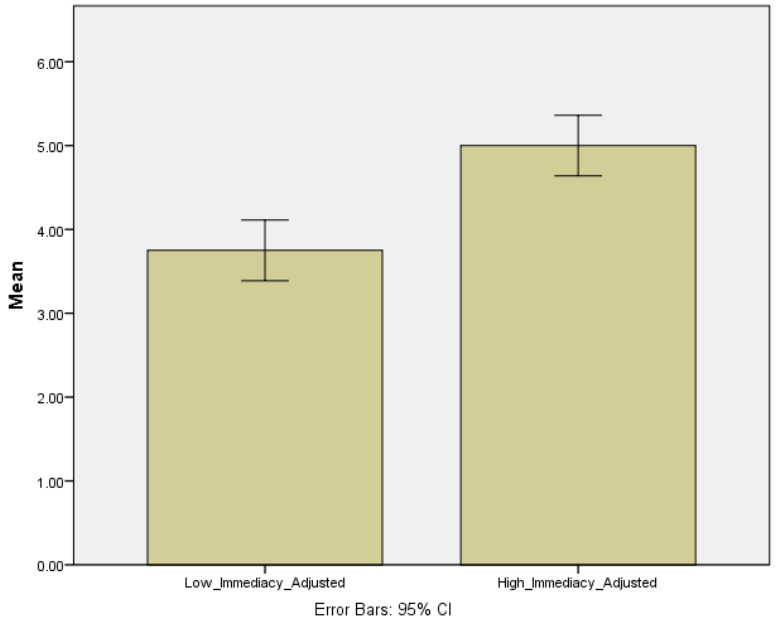
No difference between two graphs because SPSS assumes scores are independent, and consequently the error bars do not reflect the 'true' error around the means for repeated measures designs

Adjusted means for within-subjects study

Confidence Intervals Generated from between-subjects study



Confidence Intervals Generated from within-subjects study with adjusted means



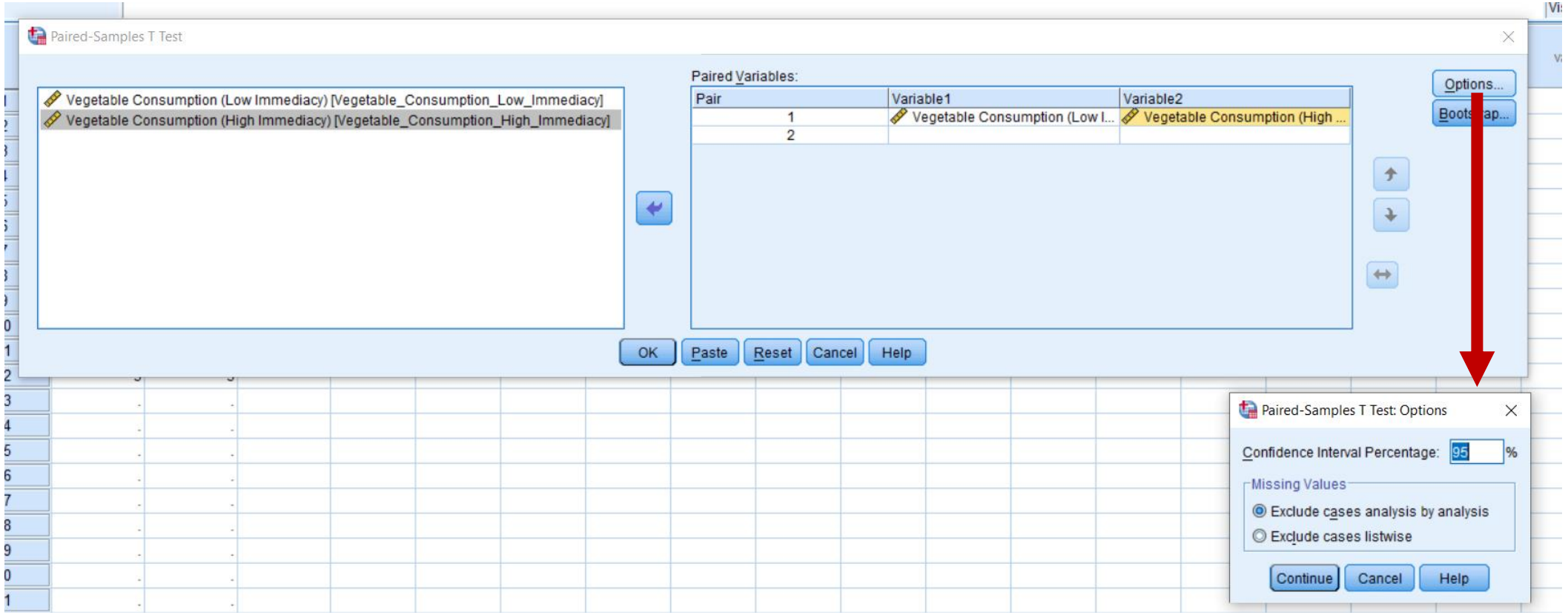
Is there a difference now?

1) Means in the two conditions have not changed. However, the error bars have got smaller.

2) Error bars previously overlapped but in this new graph they do not.

3) The plots of the error bars shows the repeated-measures data is extra sensitive to differences

Paired-samples *t*-test using SPSS Statistics



Paired- samples *t*-test output

Paired Samples Statistics

				Bootstrap[a]			
						BCa 95% Confidence Interval	
			Statistic	Bias	Std. Error	Lower	Upper
Pair 1	Vegetable Consumption (Low Immediacy)	Mean	3.75	.02	.54	2.50	4.92
		N	12				
		Std. Deviation	1.913	-.129	.352	1.357	2.193
		Std. Error Mean	.552				
	Vegetable Consumption (High Immediacy)	Mean	5.00	.02	.47	3.83	6.17
		N	12				
		Std. Deviation	1.651	-.103	.310	1.115	1.946
		Std. Error Mean	.477				

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Paired Samples Correlations

					Bootstrap for Correlation[a]			
							BCa 95% Confidence Interval	
		N	Correlation	Sig.	Bias	Std. Error	Lower	Upper
Pair 1	Vegetable Consumption (Low Immediacy) & Vegetable Consumption (High Immediacy)	12	.806	.002	-.029	.160	.417	.943

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Paired-samples *t*-test output

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Vegetable Consumption (Low Immediacy) - Vegetable Consumption (High Immediacy)	-1.250	1.138	.329	-1.973	-.527	-3.804	11	.003

Bootstrap for Paired Samples Test

		Mean	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
Pair 1	Vegetable Consumption (Low Immediacy) - Vegetable Consumption (High Immediacy)	-1.250	-.005	.311	.008	-1.833	-.667

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Calculating an effect size

$$r = \sqrt{\frac{(-3.804)^2}{(-3.804)^2 + 11}} = \sqrt{\frac{14.47}{25.47}} = .75$$

$$\hat{d} = \frac{(\bar{X}_{High\ Immediacy} - \bar{X}_{Low\ Immediacy})}{s_{Low\ Immediacy}} = \frac{5 - 3.75}{1.91} = 0.65$$

Reporting the paired-samples *t*-test

On average, participants given a High Immediacy Robot ate vegetables more frequently (M = 5, SE = 0.48), than those given a Low Immediacy Robot (M = 3.75, SE = 0.55). This difference, –1.25, BCa 95% CI [−1.833, −0.667], was significant $t(11) = -3.80$, p = .003 and represented a medium-sized effect d = .65.

Drawing Conclusions

- It is often assumed that when *p* is less than 0.05 it can be assumed that the independent variable had an effect and it was not purely chance
- However, there are cases where it is mistakenly concluded that an independent variable had an effect

Type I error – Independent variable is mistakenly concluded as having an effect.

Why is this bad?

- You would end up developing false theories that have no impact

How often will you get a Type I error with a $p < 0.05$?

Comparing several independent Conditions

Comparing several means with the linear model

- Theory
- Implementation using SPSS Statistics

Testing specific hypotheses:

- Planned Contrasts/Comparisons
  - Choosing Contrasts
  - Coding Contrasts
- Post Hoc Tests

A Robot-Assisted Therapy Example

A robot therapy Randomized Control Trial (RCT) in which we randomized people into three groups:

1. A control group
2. 15 minutes of robot therapy
3. 30 minutes of robot contact.

The outcome is happiness (0 = unhappy) to 10 (happy).

Predictions:

1. Any form of robot therapy should be better than the control (i.e. higher happiness scores)
2. A dose-response hypothesis that as exposure time increases (from 15 to 30 minutes) happiness will increase too.

The data

$\overline{X}$	Control	15 minutes	30 minutes
	3	5	7
	2	2	4
	1	4	5
	1	2	3
	4	3	6
	2.20	3.20	5.00
s	1.30	1.30	1.58
s^2	1.70	1.70	2.50
Grand mean = 3.467 Grand SD = 1.767 Grand variance = 3.124			

The linear model

$$\text{Happiness}_i = b_0 + b_1\text{Long}_i + b_2\text{Short}_i + \varepsilon_i$$

Group	Dummy Variable 1 (Long)	Dummy Variable 2 (Short)
Control	0	0
15 minutes of robot therapy	0	1
30 minutes of robot therapy	1	0

Control group

$$\text{Happiness}_i = b_0 + (b_1 \times 0) + (b_2 \times 0)$$
$$\text{Happiness}_i = b_0$$
$$\bar{X}_{\text{control}} = b_0$$

30-minute group

$$\text{Happiness}_i = b_0 + (b_1 \times 1) + (b_2 \times 0)$$

$$\text{Happiness}_i = b_0 + b_1$$

$$\bar{X}_{30\ \text{mins}} = \bar{X}_{\text{control}} + b_1$$

$$b_1 = \bar{X}_{30\ \text{mins}} - \bar{X}_{\text{control}}$$

15-minute group

$$\text{Happiness}_i = b_0 + (b_1 \times 0) + (b_2 \times 1)$$

$$\text{Happiness}_i = b_0 + b_2$$

$$\bar{X}_{15\text{ mins}} = \bar{X}_{\text{control}} + b_2$$

$$b_2 = \bar{X}_{15\text{ mins}} - \bar{X}_{\text{control}}$$

Using the regression menus

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20.133	2	10.067	5.119	.025[b]
	Residual	23.600	12	1.967		
	Total	43.733	14			

a. Dependent Variable: Happiness (0-10)

b. Predictors: (Constant), Dummy 2: 15 mins vs. Control, Dummy 1: 30 mins vs. Control

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.200	.627		3.508	.004
	Dummy 1: 30 mins vs. Control	2.800	.887	.773	3.157	.008
	Dummy 2: 15 mins vs. Control	1.000	.887	.276	1.127	.282

a. Dependent Variable: Happiness (0-10)

Experiments vs. natural observation

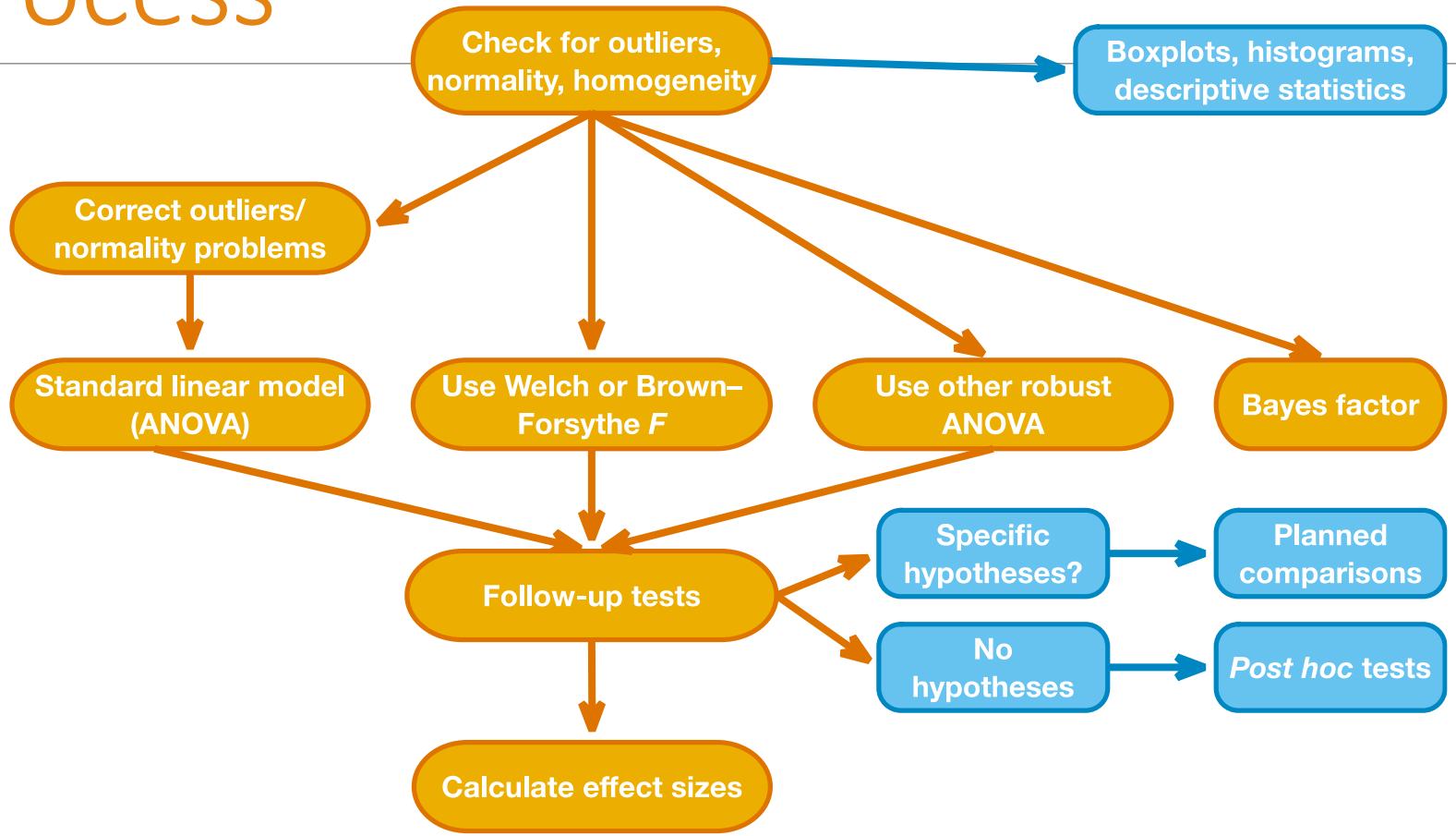
Natural observation

- Differences between means are not (necessarily) caused by the predictor variable being measured

Experiments

- The predictor variable is manipulated in a controlled way, usually to compare conditions in which a causal variable is present to conditions where it is absent
- In such cases causal inferences can be made

The process


Check for outliers, normality, homogeneity
Boxplots, histograms, descriptive statistics
Correct outliers/ normality problems
Standard linear model (ANOVA)
Use Welch or Brown–Forsythe *F*
Use other robust ANOVA
Bayes factor
Follow-up tests
Specific hypotheses?
Planned comparisons
No hypotheses
Post hoc tests
Calculate effect sizes

Theory of the *F*-statistic

We calculate how much variability there is between scores

- Total Sum of squares (SS_T).

We then calculate how much of this variability can be explained by the model we fit to the data

- How much variability is due to the predictor variable/experimental manipulation, Model Sum of Squares (SS_M)...

… and how much cannot be explained

- How much variability is due to individual differences in performance, Residual Sum of Squares (SS_R).

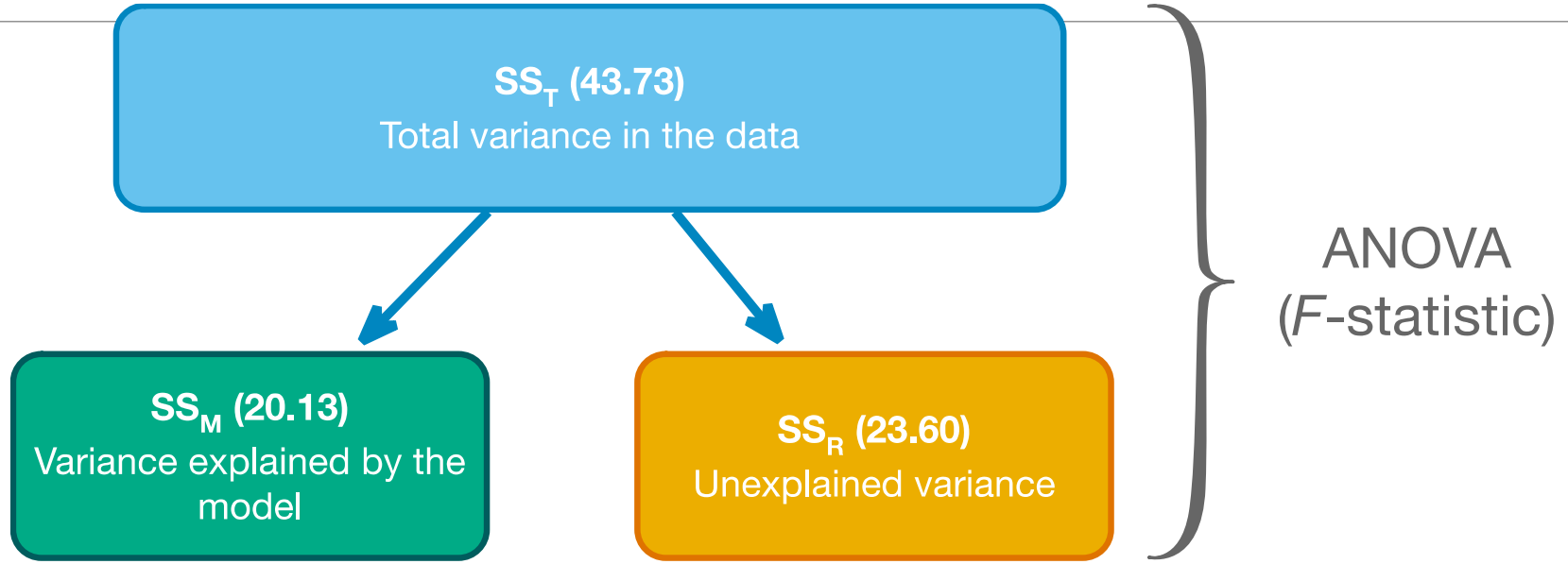
Theory of the *F*-statistic

We compare the amount of variability explained by the model (experiment), to the error in the model (individual differences)

- This ratio is called the *F*-statistic

If the model explains a lot more variability than it can't explain, then the experimental manipulation has had a significant effect on the outcome (DV).

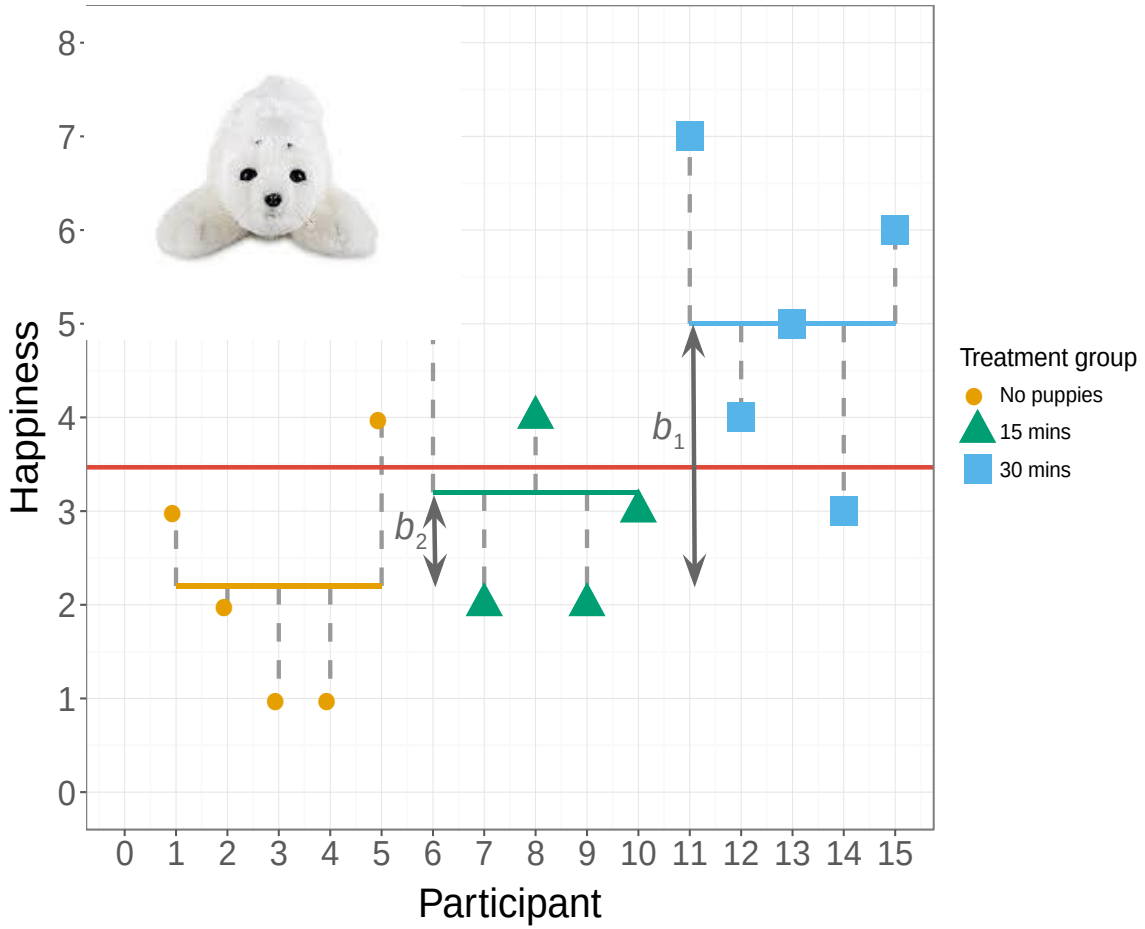
Theory of the *F*-statistic



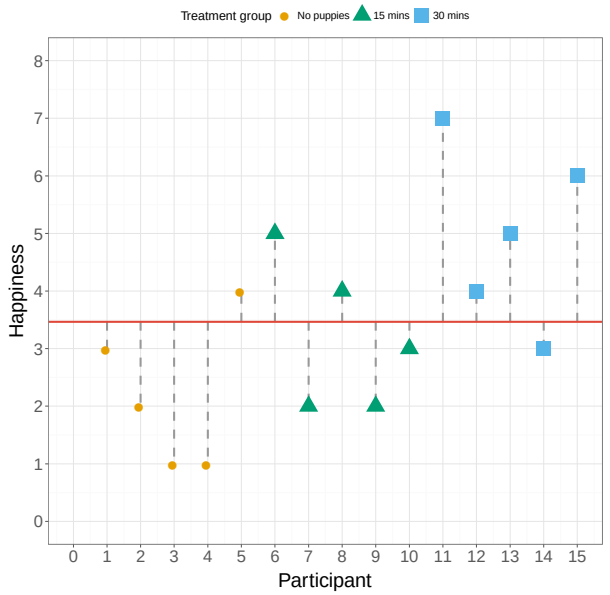
If the experiment is successful, then the model will explain more variance than it can't

- SS_M will be greater than SS_R

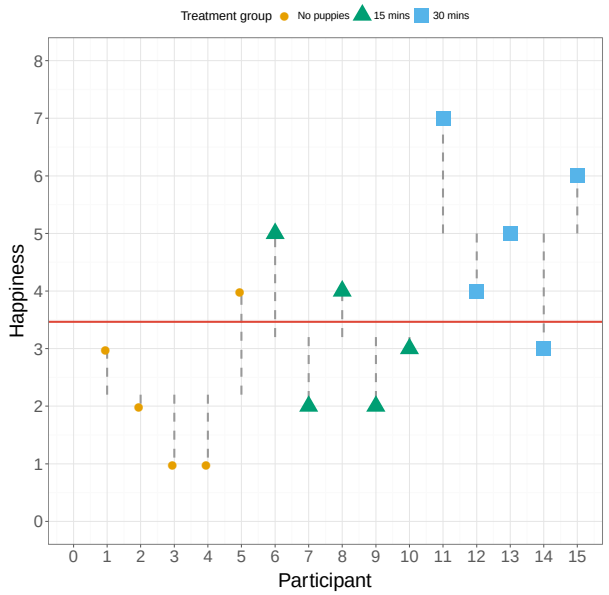
The model:

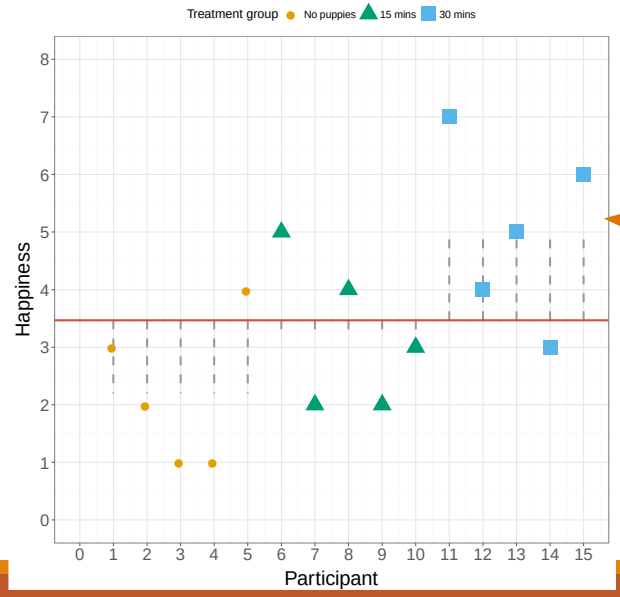

Happiness
Participant
Treatment group
No puppies
15 mins
30 mins
b_1
b_2

SS_T uses the differences between the observed data and the mean value of Y



SS_R uses the differences between the observed data and the model (group means)





SS_M uses the differences between the mean value of Y and the model (group means)

Step 1: Calculate SS_T

$$SS_T = \sum_{i=1}^{N} \left(x_i - \bar{x}_{\text{grand}}\right)^2$$

$$\begin{aligned} SS_T &= s^2_{\text{grand}}(N-1) \\ &= 3.124(15-1) \\ &= 3.124 \times 14 \\ &= 43.74 \end{aligned}$$

Degrees of freedom (*df*)

Degrees of Freedom (*df*) are the number of values that are free to vary.

- Think about soccer teams!

In general, the *df* are one less than the number of values used to calculate the SS.

$$df_{\mathrm{T}} = N - 1 = 14$$

Step 2: Calculate SS_M

$$SS_M = \sum_{g=1}^{k} n_g \left(\bar{x}_g - \bar{x}_{\text{grand}}\right)^2$$

$$SS_M = 5(2.200 - 3.467)^2 + 5(3.200 - 3.467)^2 + 5(5.000 - 3.467)^2$$
$$SS_M = 8.025 + 0.355 + 11.755$$
$$SS_M = 20.135$$

Model degrees of freedom

How many values did we use to calculate SS_M?

- We used the 3 means.

$$df_M = k - 1 = 2$$

Step 3: Calculate SS_R

$$SS_R = \sum_{g=1}^{k} \sum_{i=1}^{n} \left(x_{ig} - \bar{x}_g\right)^2$$

$$SS_R = \sum_{g=1}^{k} s_g^2 \left(n_g - 1\right)$$

Step 3: Calculate SS_R

$$SS_R = s^2_{\text{group } 1}(n_1 - 1) + s^2_{\text{group } 2}(n_2 - 1) + s^2_{\text{group } 3}(n_3 - 1)$$
$$SS_R = 1.70(5 - 1) + 1.70(5 - 1) + 2.50(5 - 1)$$
$$SS_R = (1.70 \times 4) + (1.70 \times 4) + (2.50 \times 4) = 6.8 + 6.8 + 10$$
$$SS_R = 23.60$$

Residual degrees of freedom

How many values did we use to calculate SS_R?

- We used the 5 scores for each of the SS for each group.

$$df_R = df_{\text{group 1}} + df_{\text{group 2}} + df_{\text{group 3}}$$
$$SS_R = (n-1) + (n-1) + (n-1)$$
$$SS_R = (5-1) + (5-1) + (5-1)$$
$$SS_R = 12$$

Step 4: Calculate the mean squared error

$$\text{MS}_M = \frac{SS_M}{df_M} = \frac{20.135}{2} = 10.067$$

$$\text{MS}_R = \frac{SS_R}{df_R} = \frac{23.60}{12} = 1.96$$

Step 5: Calculate the *F*-statistic

$$F = \frac{MS_M}{MS_R}$$

$$F = \frac{MS_M}{MS_R} = \frac{10.067}{1.967} = 5.12$$

Step 6: construct a summary table

Source	SS	df	MS	*F*
Model	20.14	2	10.067	5.12*
Residual	23.60	12	1.967	
Total	43.74	14		

Robust *F*

Report and interpret the robust versions of *F* by default

Robust Tests of Equality of Means

Happiness (0-10)

	Statistic[a]	df1	df2	Sig.
Welch	4.320	2	7.943	.054
Brown-Forsythe	5.119	2	11.574	.026

a. Asymptotically F distributed.

Why use follow-up tests?

The *F*-statistic tells us only that the experiment was successful

- i.e. group means were different

It does not tell us specifically which group means differ from which.

We need additional tests to find out where the group differences lie.

How?

Orthogonal Contrasts/Comparisons

- Hypothesis driven
- Planned a priori

Post Hoc Tests

- Not Planned (no hypothesis)
- Compare all pairs of means

Trend Analysis

Planned contrasts

Basic Idea:

- The variability explained by the Model (experimental manipulation, **SS_M)** is due to participants being assigned to different groups.
- This variability can be broken down further to test specific hypotheses about which groups might differ.
- We break down the variance according to hypotheses made *a priori* (before the experiment).
- It's like cutting up a cake (yum yum!)

Rules when choosing contrasts

Independent

- contrasts must not interfere with each other (they must test unique hypotheses).

Only 2 Chunks

- Each contrast should compare only 2 chunks of variation (why?).

K-1

- You should always end up with one less contrast than the number of groups.

How do I Choose Contrasts?

Big Hint:

- In most experiments we usually have one or more control groups.
- The logic of control groups dictates that we expect them to be different to groups that we've manipulated.
- The first contrast will always be to compare any control groups (chunk 1) with any experimental conditions (chunk 2).

Hypotheses

Hypothesis 1:

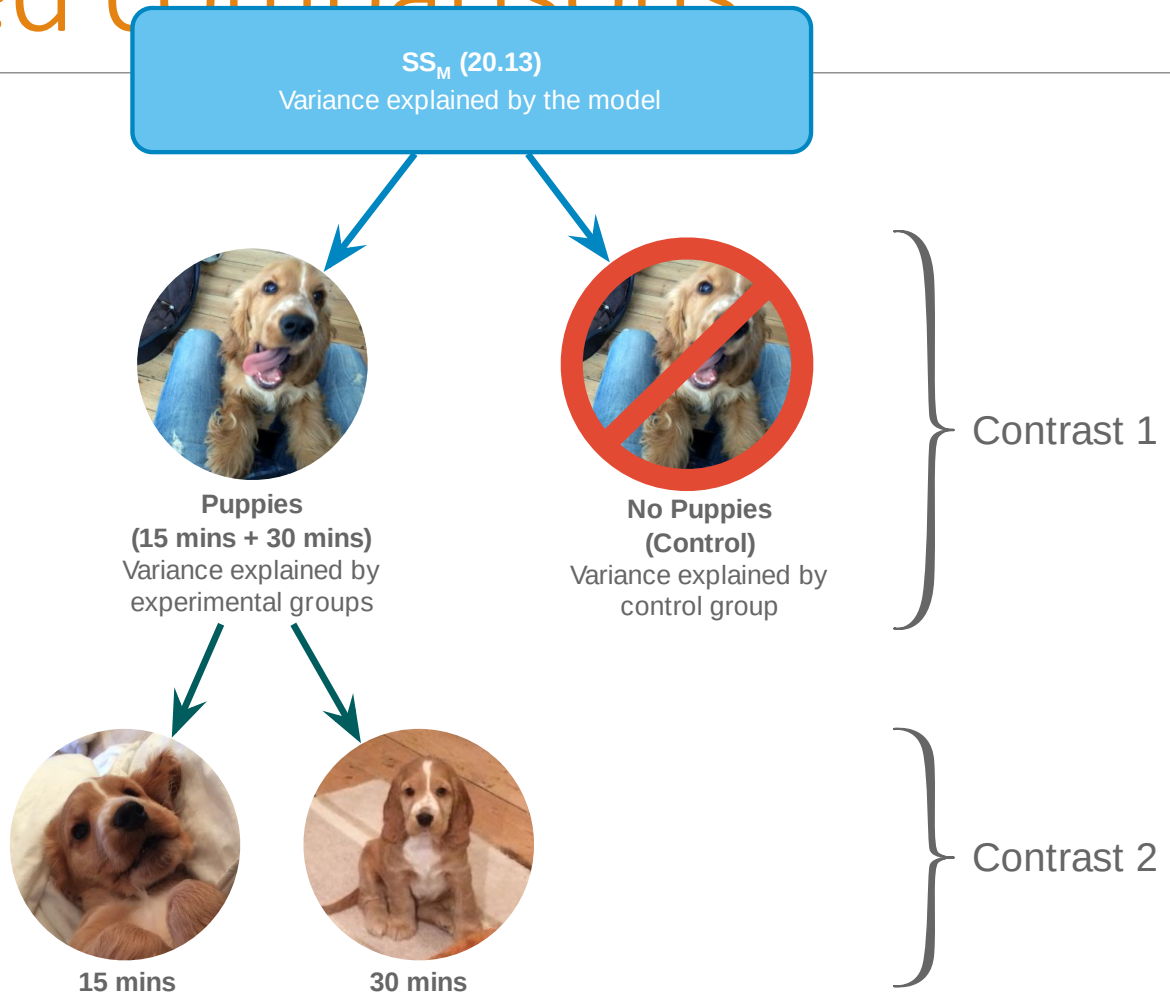
- People who receive robot therapy will be happier than those who don’t.
- Control ≠ (15 mins, 30 mins)

Hypothesis 2:

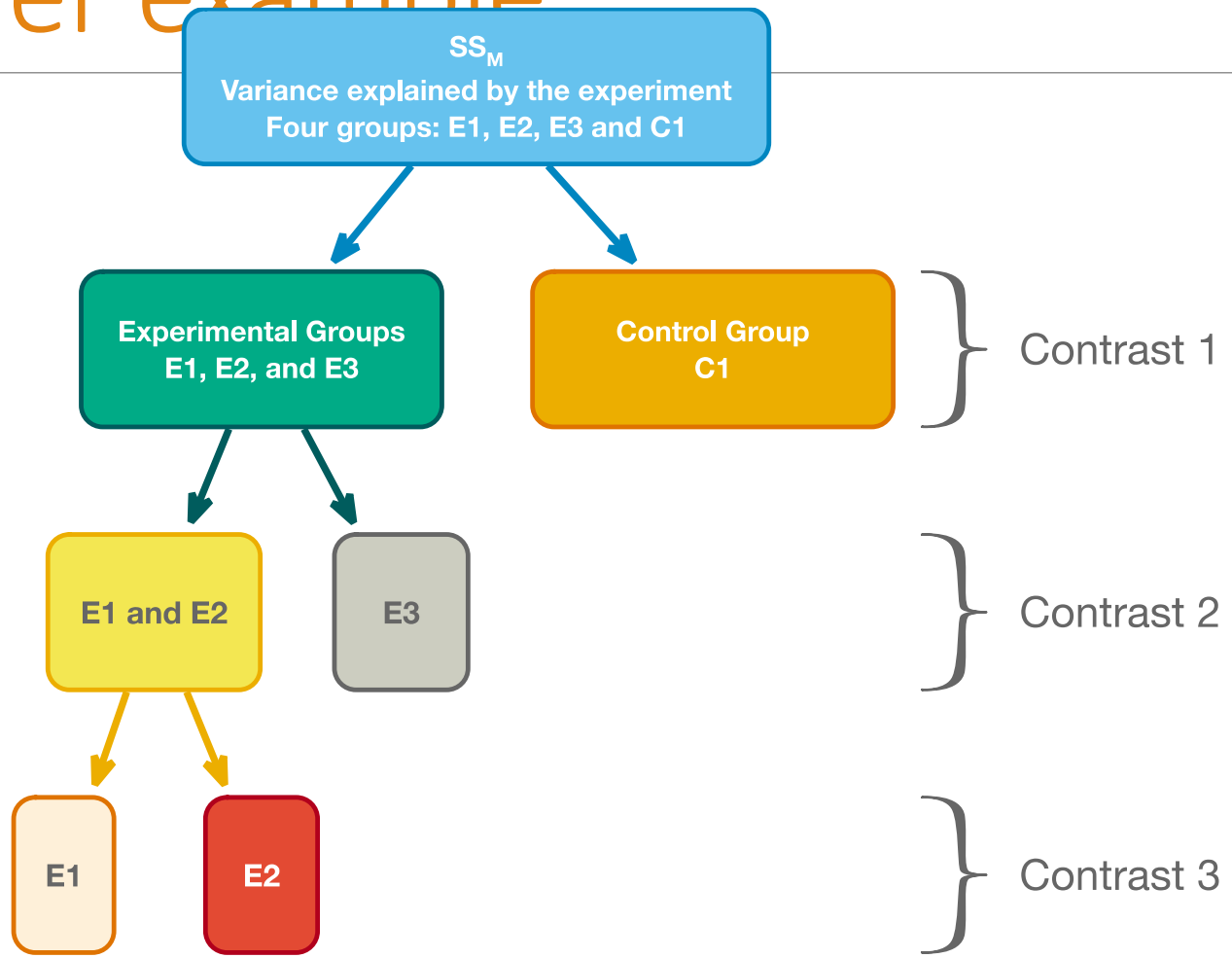
- People who receive a long dose of robot therapy will be happier than those who receive a shortdose
- 15 mins ≠ 30 mins

	Control	15 mins	30 mins
	3	5	7
	2	2	4
	1	4	5
	1	2	3
	4	3	6
Mean	2.20	3.20	5.00

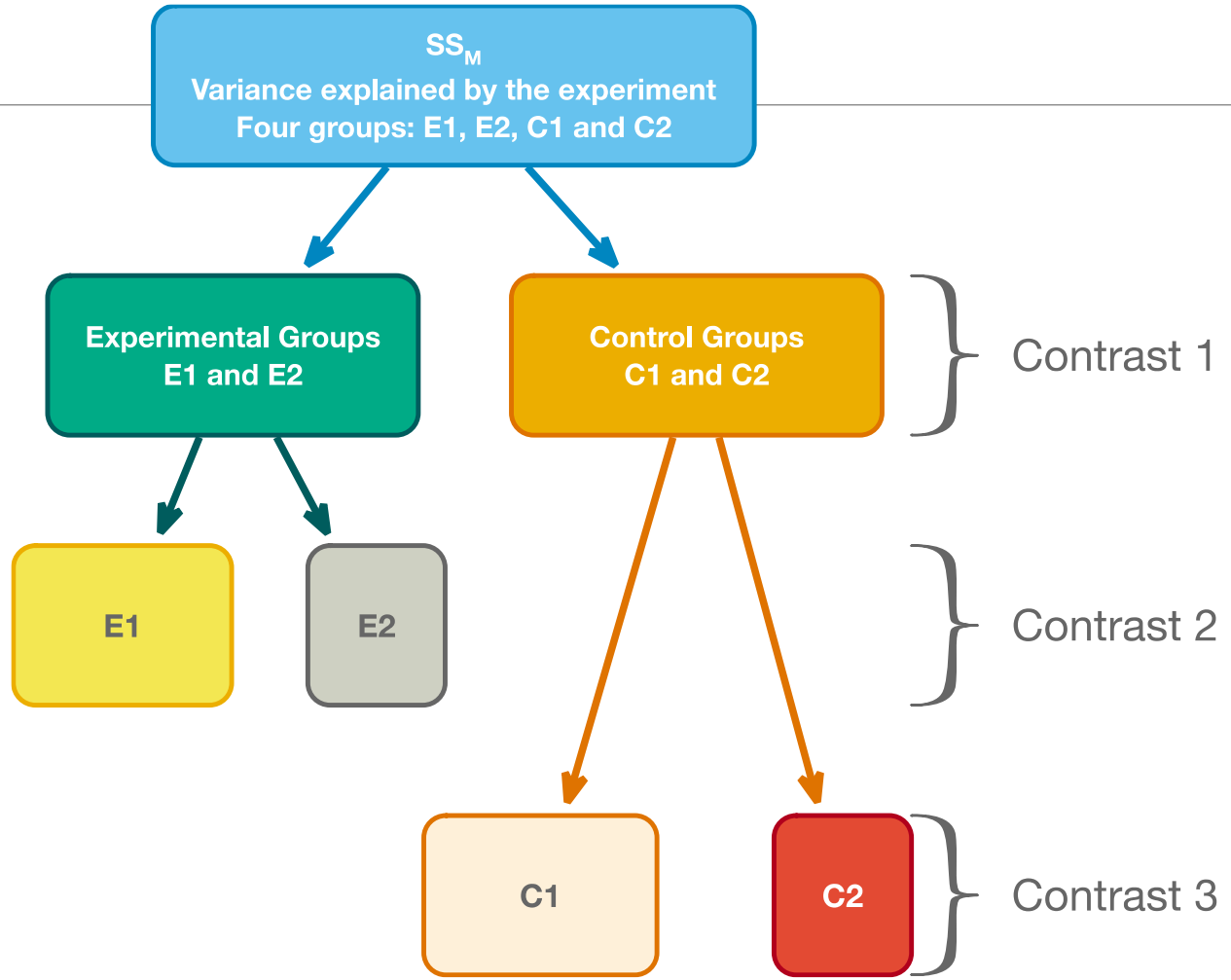
Planned comparisons



Another example



Another example



Coding planned contrasts: rules

Rule 1

- Groups coded with positive weights compared to groups coded with negative weights

Rule 2

- The sum of weights for a comparison should be zero

Rule 3

- If a group is not involved in a comparison, assign it a weight of zero

Coding planned contrasts: rules

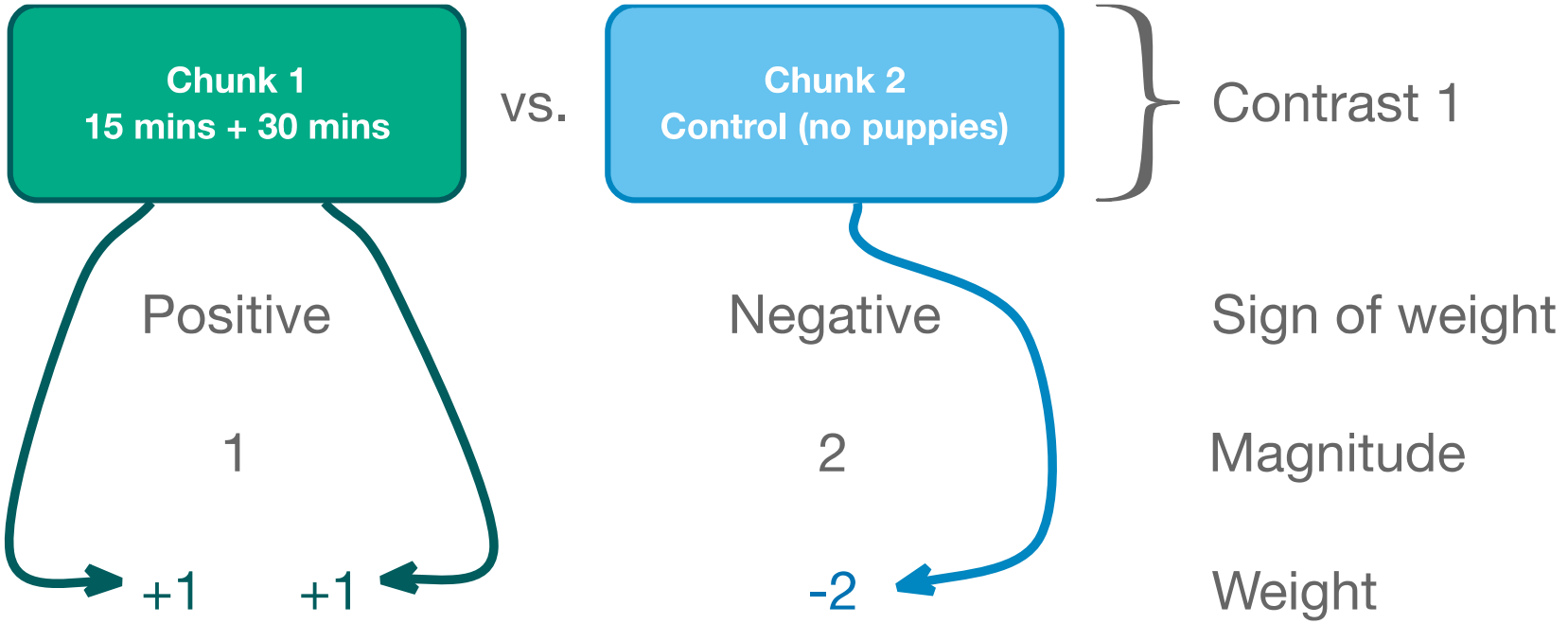
Rule 4

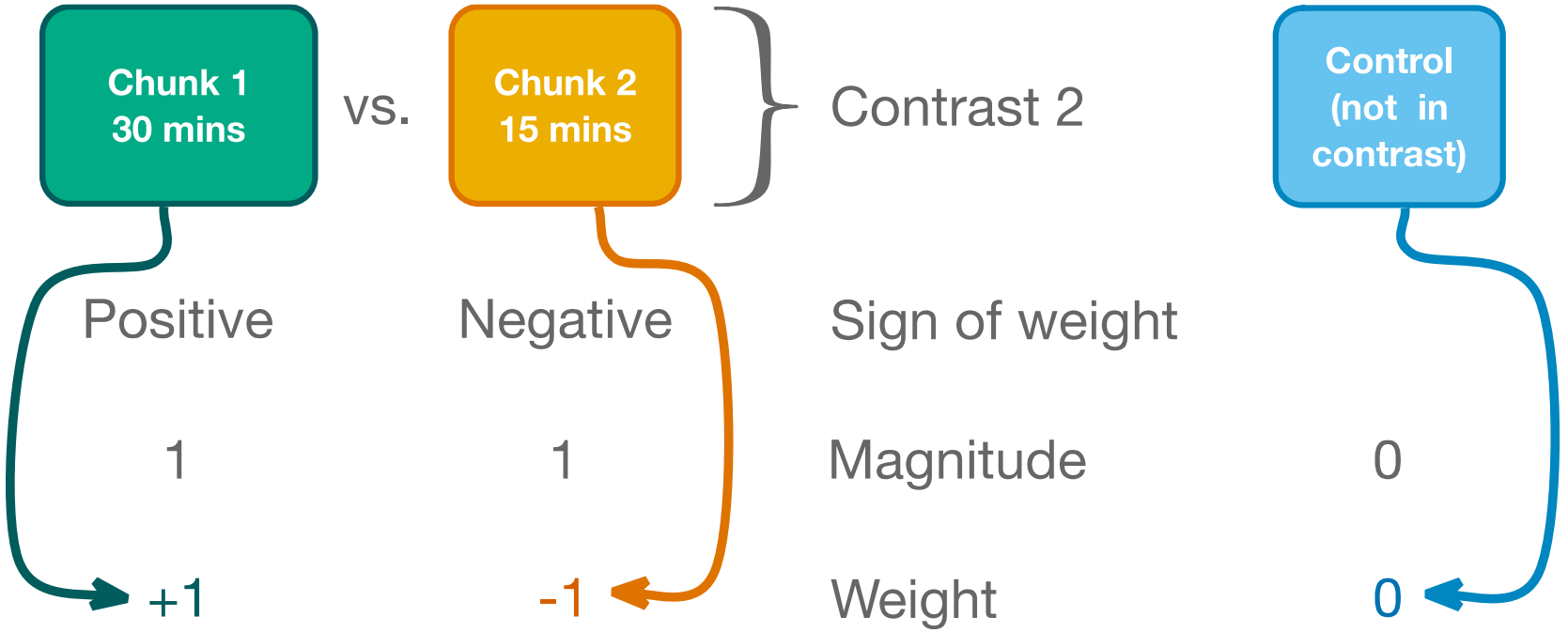
- For a given contrast, the weights assigned to the group(s) in one chunk of variation should be equal to the number of groups in the opposite chunk of variation

Rule 5

- If a group is singled out in a comparison, then that group should not be used in any subsequent contrasts

Defining contrasts using weights




Chunk 1
30 mins
vs.
Chunk 2
15 mins
Contrast 2
Control
(not in
contrast)
Positive
Negative
Sign of weight
1
1
Magnitude
0
+1
-1
Weight
0

Output

Contrast Coefficients

	Dose of puppies		
Contrast	Control	15 mins	30 mins
1	-2	1	1
2	0	-1	1

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Happiness (0-10)	Assume equal variances	1	3.80	1.536	2.474	12	.029
		2	1.80	.887	2.029	12	.065
	Does not assume equal variances	1	3.80	1.483	2.562	8.740	.031
		2	1.80	.917	1.964	7.720	.086

Reporting Results Continued

Planned contrasts revealed that having any dose of robot therapy significantly increased happiness compared to having a control, $t(8.74) = 2.56$, $p = 0.031$, $r = 0.58$, but having 30 minutes did not significantly increase happiness compared to having 15 minutes, $t(7.72) = 1.96$, $p = 0.086$, $r = 0.51$.

Quasi-Experiments

- Utilized when we cannot directly manipulate the independent variable
- Similar to “experimental” research methods where we still have groups and statistical analyses.
- The most common Quasi-Experimental design is **non-equivalent control group design** where subjects are grouped by: 1) existing groups or 2) self-selection

Quasi-Experiments

Existing Groups or Ex Post Facto Design – In many experiments existing groups are utilized as the different levels or conditions of an independent variable

Example Scenario: We hypothesize that the longer a user has had the robot the less they will crash into walls during teleoperation.

- In this scenario the independent variable is the length of time a user has had the robot (<1y, 1-10y, >10y) and the dependent variable is # of crashes during teleoperation
- We cannot control the independent variable because we cannot assign how long a user has had a robot for. What's the issue with this?
  - User's that fall into specific groups may also vary along other dimensions (e.g. users that had the robot longer may be wealthier and as a result money to pay for also additional lessons)
  - Hence, we have less control over Quasi-Experiments and cannot rule out confounding factors.

Quasi-Experiments

Subject Self-Selection – In this case, we allow the subjects to select the condition they will be tested on.

- This has several design issues because we can't know whether there are potential confounds for why users are choosing a particular condition.
  - For example, let's say we were testing different modalities for robot control such as a mouse vs. keyboard and we allowed participants to choose the condition for accomplishing the task.
    - The mouse users are shown to do better then the keyboard users
    - However, what if we did not realize that self selected mouse users had more experience and utilized that modality because they were more familiar with it.
    - We then have a confounding factor of experience and the effect of modality may be misrepresented by experience
    - Hence, we need to be careful with Quasi-Experiments and biases due to confounding factors.

Descriptive Methods

- While tightly controlled experimentation is valuable for learning basic laws and principles it may not be representative of real-world environments

- Since we cannot control or manipulate variables in a real-world settings Descriptive research is utilized where a number of variables are measured and the relationships to each other are evaluated.

Descriptive Methods - Observation

The observation method consists of recording behavior during tasks while performed in specific contexts.

- In an observational study, researchers identify:
  - Variables to be measured
  - Methods for observing the variable
  - Conditions under which the observation will occur
  - Observation time frame
  - Having a checklist of well defined variables will ensure the observer stays focused on the correct details
  - This information should be defined from pilot data and not changed after starting the observations because it will bias the results as well as lead to excessive amounts of data.

- Limitations of observational studies:
  - Observations are likely to affect the behavior being measured
  - Reliability of the observer because they can change their interpretation over time. This can be mitigated by interobserver reliability methods such as having multiple observers of a behavior.
    - Cohen's Kappa and reviewing videotapes to discuss conflicting observations

Descriptive Methods – Surveys and Questionnaires

Surveys – Systematic gathering of information about people's beliefs, attitudes, values, and behavior

Questionnaire – Set of questions or scales used for experimental and descriptive research

- For experimental research questionnaires are either administered after each condition or at the end of all the conditions. Statistical analyses are then conducted to see if there are any differences
- For descriptive research questionnaires are also administered except independent variables are not manipulated.

Two types of questionnaires: 1) self-administered and 2) interviewer-administered

- Self-administered are more common as they are faster and efficient

Descriptive Methods – Surveys and Questionnaires

Content

1) The content of the questionnaire is developed according to the dependent variables of interest. (e.g. ease of use of the buttons, joystick, and screen)
2) These variables can then be grouped into categories. (e.g. ease of use)
3) The questionnaire items can then be specifically defined using clear and concise questions

Format

1) Open-ended (e.g. what do you like about the controller). Good for when you cannot forsee all possible answers, range of answers is too large, or when researcher wants respondent's own words
2) Closed question format where respondents are limited to their answers. This is often easier to get quantitative results which can be easily used for inferential or descriptive statistics.

How hard is it to use the controller to navigate a robot							
Extremely Difficult	1	2	3	4	5	6	7	Extremely Easy

Descriptive Methods – Surveys and Questionnaires

Pretesting – After questionnaire is written up it should be pretested to see if there are any potential issues with the questions. Statistical analyses is not required at this step

- There is not enough resolution with the question responses
- Questions may not answer what the researcher desires

Validity – Questionnaires are only as valid to the extent that they are well designed and measure what they intend to. Questionnaires often measure beliefs and behavior that are sensitive or open to judgement

- Responses should be made confidential and anonymous to ensure respondents can answer honestly
- Deidentifying users with numbers helps with this process

Data Analysis for Descriptive Measures

The primary goals of descriptive measures whether observational or questionnaires are to:

1) See whether relationships exist
2) Measure the strengths of these relationships

Differences between Groups - Evaluate how dependent variables differ for different groups (i.e. independent variables) using ANOVAs

Relationships between Continuous Variables – We can also measure if there is a relationship between a dependent variable without grouping subjects.

- This is done via a correlational analysis which identifies the extent to which two variables covary.
- These statistical tests will then provide information on whether a relationship exists (*p*) and the strength of the relationship (*r*).
- Caution should be noted when finding statistically significant correlation
  - Causation could be in the opposite direction where the dependent variable is causing the independent variable
  - A third variable could be causing changes in both the dependent and independent variables.

Measuring Variables

1) Subjective measures
2) Performance
3) Physiological

Qualities of Good Research

Construct Validity – This refers to the degree which researchers manipulated what they wanted to and measured the dependent variables they aimed to.

- If we were looking at the independent variable “fatigue” where fatigue refers to the state of excessive physical effort then subjects need to be truly fatigued
  - For example, in a driving study on drowsiness manipulation checks need to be done to ensure that the study participant is actually made drowsy
- Similarly, the dependent measure could be measured incorrectly.
  - For example, for measuring user interface “efficiency” it may be more important to reduce the “number of actions” required to complete a task then it is to reduce the “time” to completion. Efficiency could also be defined by frequency of “useless/unnecessary actions”

Internal Validity– High internal validity refers to when casual or independent variables are the only causes in change in the dependent variables being measured.

External Validity – This refers to how the results can generalize to other people, tasks, and/or settings.

Ethical Issues

- Code of Federal Regulations Protections of Human Subjects
- APA guidelines for ethical treatment of human subjects which has the following principles:
  - Protection of participants from mental or physical harm
  - The right of participants to privacy with respect to their behavior
  - The assurance that participation in research is completely voluntary
  - The right of participants to be informed beforehand about the nature of the experimental procedures
- Institutional Review Board (IRBs)
  - Most hospitals, universities, and other organizations usually have their own IRBs responsible for ensuring proper conduct of research.
  - Research studies are typically described and documented in detail so that they can be submitted to these IRBs.
  - A panel of impartial experts are then responsible to ensure that the methods proposed for the research are ethical.

Ethical Issues

General guidelines :

- Participants should be told the general nature of the study but the exact hypotheses cannot be described to prevent bias of behavior
- Participant should be told their information and results should be kept anonymous and confidential
- Informed consent form should be signed stating that they understand the nature and risk of the experiment, their participation is voluntary, and they may withdraw at any time
- Reasonable risk is considered anything that is no greater than those faced in their daily lives.
- Participants should always be treated with respect and should not feel their performance is being evaluated or fear that they are not doing well. This is why the term “user testing” has been changed to “usability testing”