Human-Computer Interaction

Step-by-step Experimental Design

Professor Bilge Mutlu

Today's Agenda

» Topic overview: *Experimental Research: Step-by-step Design Guide*

» Hands-on Activity: *Experimental Design Choices for Projects*

What are the steps involved in designing an experiment?

1. **Step 1:** Formulate research question
2. **Step 2:** Identify variables
3. **Step 3:** Generate hypotheses
4. **Step 4:** Determine experimental design
5. **Step 5:** Develop experimental task & procedure
6. **Step 6:** Determine manipulations & measurements
7. **Step 7:** Identify participants

hypothesis-testing

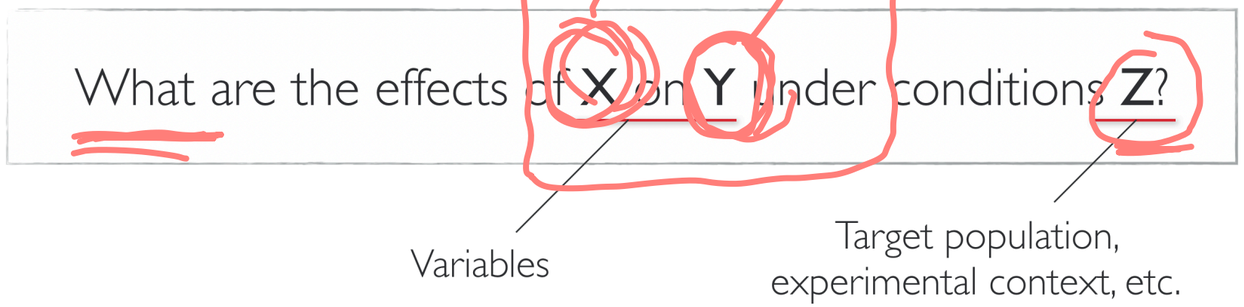
Step 1: Formulate research question

What is a research question?

Definition: The central issue to be resolved by a formal dissertation, thesis, or research project.[1]

» Should be specific enough and identify variables of interest.

» Should express the conditions under which the experiment will be performed (e.g., target population, experimental context).



[1] Duignan, 2016, Research question

How do I make sure that my research question is good?

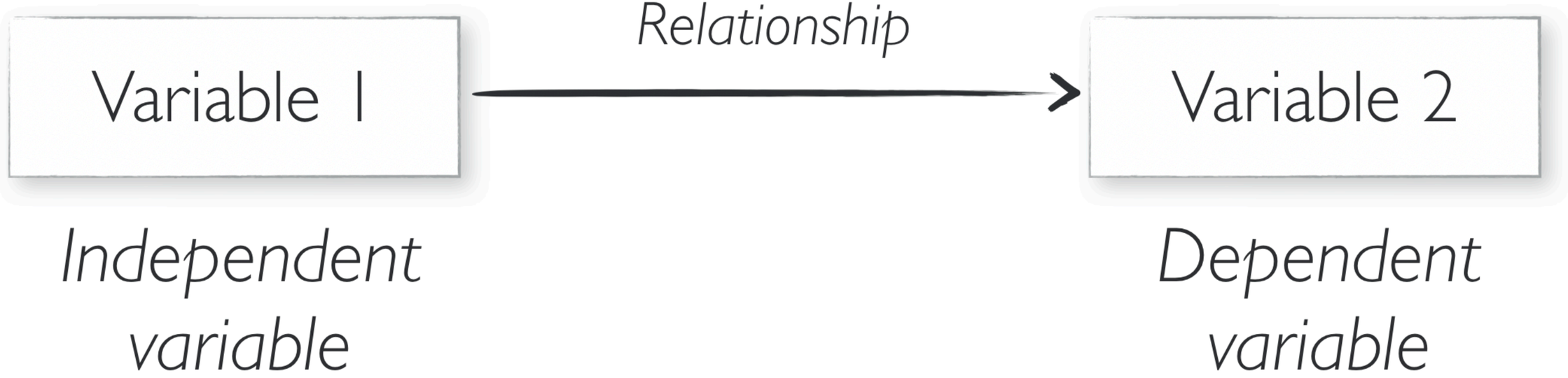
FINER criteria for good research questions:[2]

- **F (Feasible)**: Adequate number of subjects, adequate technical expertise, affordable in time and money, manageable in scope
- **I (Interesting)**: The answer intrigues investigator, peers, community
- **N (Novel)**: Confirms, refutes, or extends previous findings
- **E (Ethical)**: Amenable to a study that the IRB will approve
- **R (Relevant)**: To science, future research, technology design

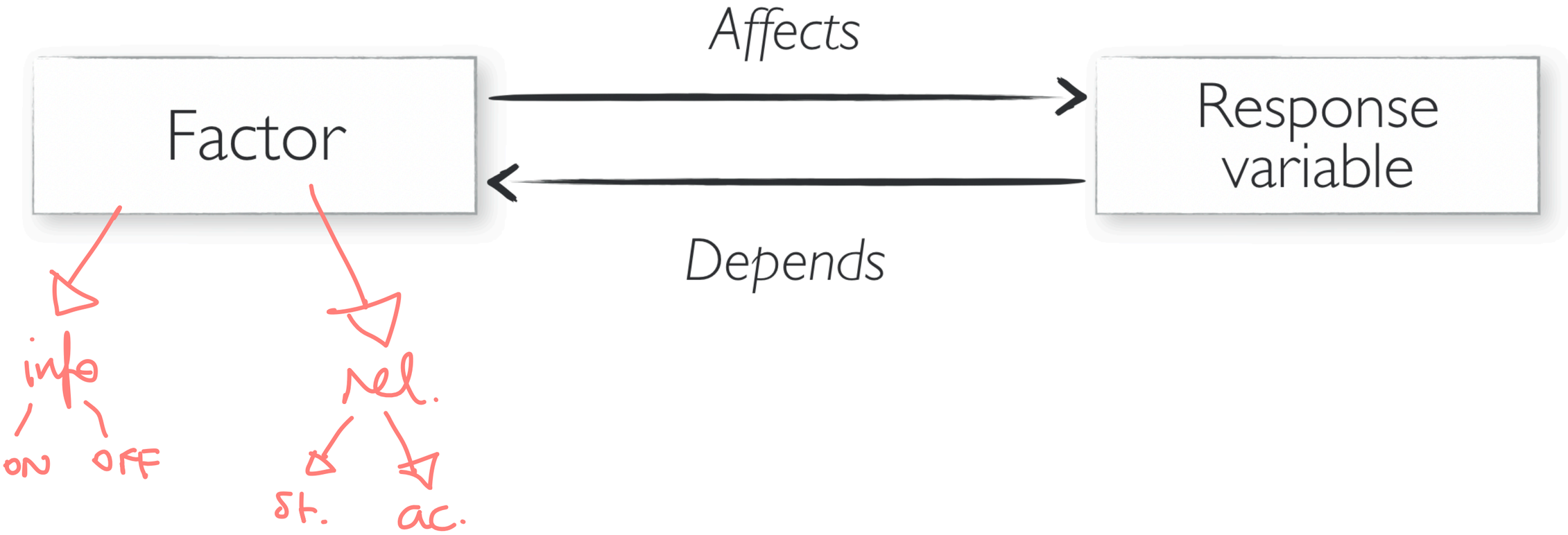
[2] Farrugia et al., 2010, Research questions, hypotheses, and objectives

Step 2: Identify variables

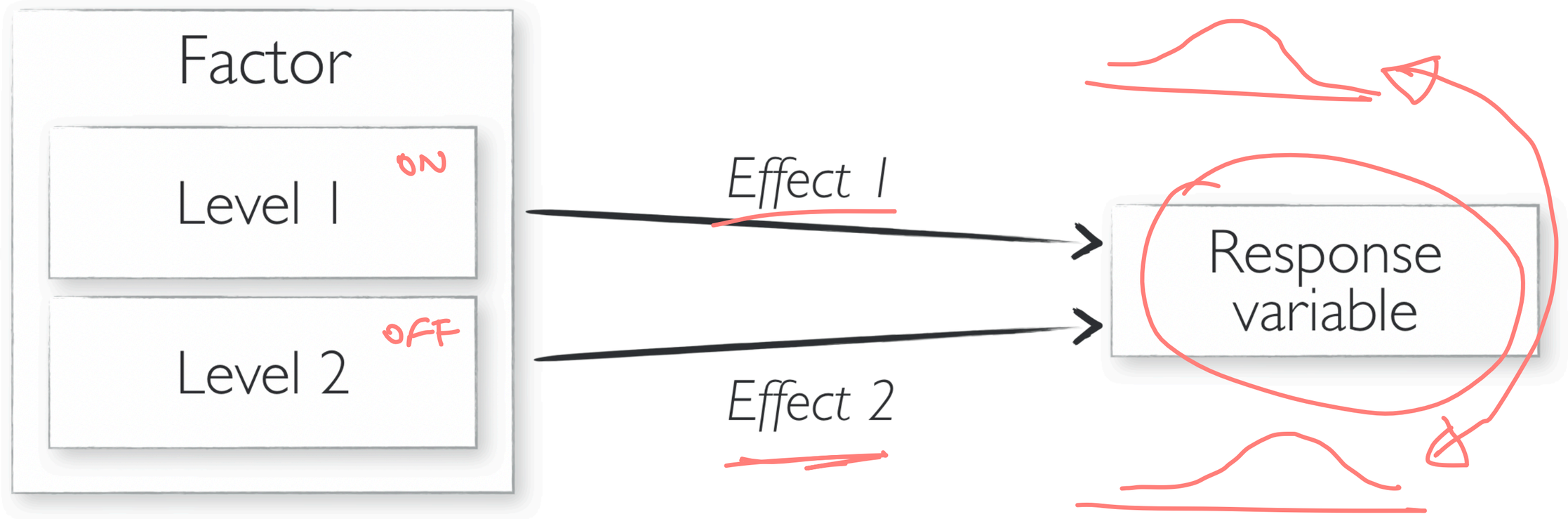
*Recap: variables of interest are **independent** and **dependent** variables that have a particular relationship; we are usually investigating the effects of independent variables on dependent variables*



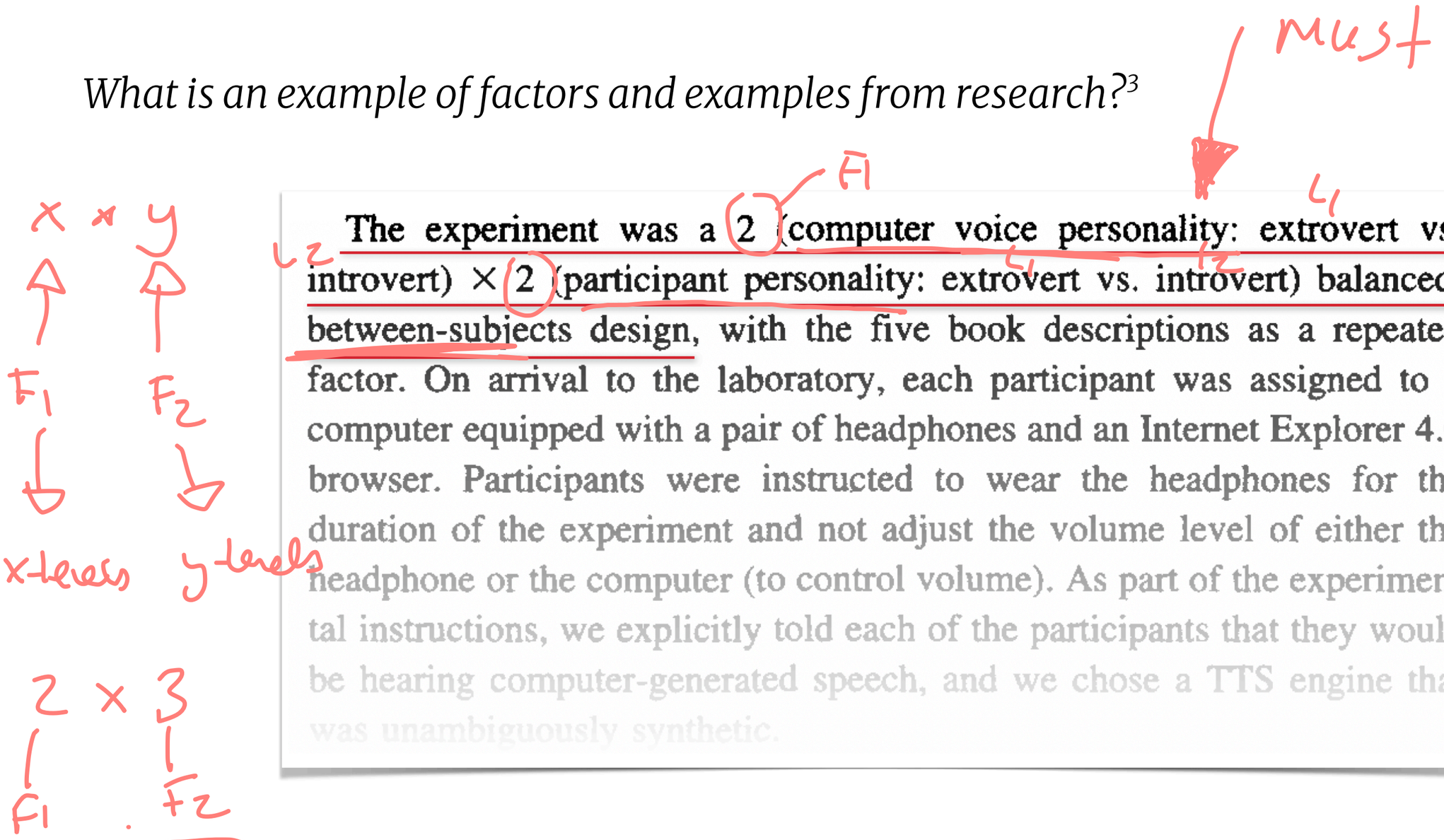
*Recap: independent variables are also called **factors**; factorial designs have at least two factors; e.g., a* **2 × 2** *design:* `caller information` *(on, off)* × `relationship` *(*`acquaintance`*,* `stranger`*)*



Recap: ***levels****, also called* ***treatment****, are the values that factors can take; e.g.,* `caller information` *can take the values* `on`*,* `off`*;* `relationship` *can take the values* `acquaintance`*,* `stranger`



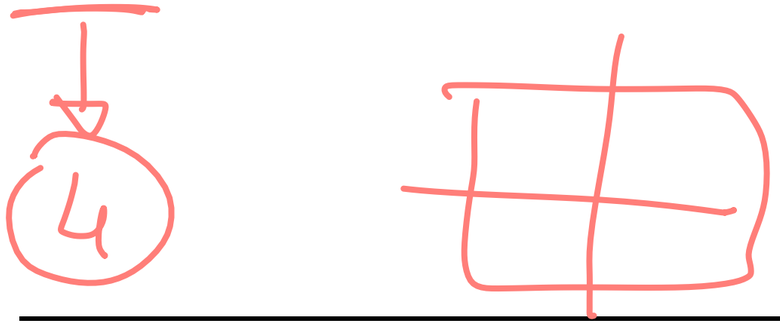
What is an example of factors and examples from research?[3]



The experiment was a 2 (computer voice personality: extrovert vs. introvert) × 2 (participant personality: extrovert vs. introvert) balanced, between-subjects design, with the five book descriptions as a repeated factor. On arrival to the laboratory, each participant was assigned to a computer equipped with a pair of headphones and an Internet Explorer 4.0 browser. Participants were instructed to wear the headphones for the duration of the experiment and not adjust the volume level of either the headphone or the computer (to control volume). As part of the experimental instructions, we explicitly told each of the participants that they would be hearing computer-generated speech, and we chose a TTS engine that was unambiguously synthetic.

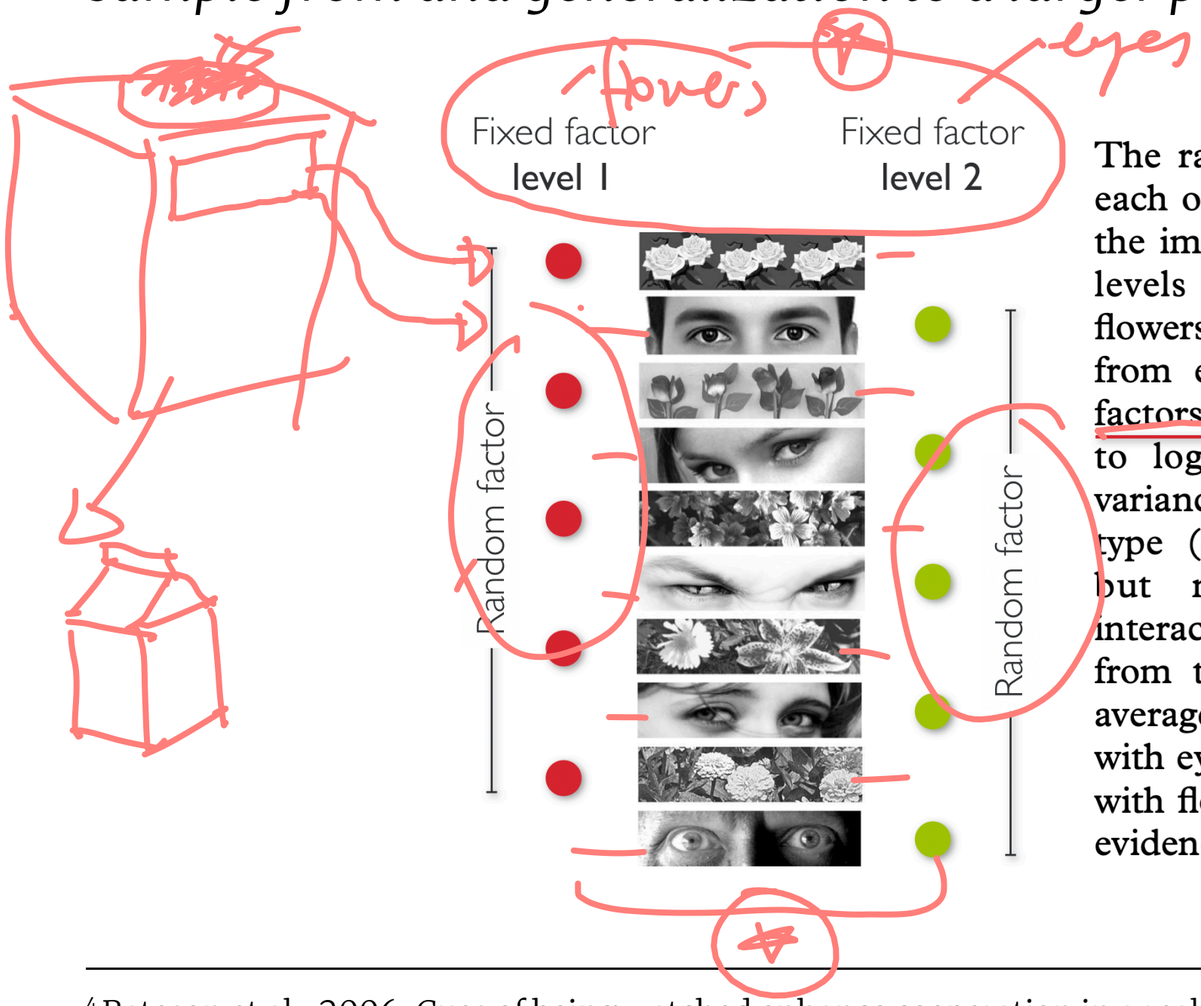
[3] Nass & Lee, 2001, Does computer-synthesized speech manifest personality?

2×2 *design with* **2** *factors,* **2** *levels each*



		Factor 1:	**Computer voice personality**
		Level 1: Extrovert	**Level 2:** Introvert
Factor 2:	**Level 1:** Extrovert	Population 1	Population 2
Participant personality	**Level 2:** Introvert	Population 3	Population 4

Recap: ***fixed factors*** *are IVs that are being studies;* ***random factors*** *that ensure a random sample from and generalization to a larger population.*[4]



The ratio of money collected to milk consumed for each of the 10 weeks is shown in figure 1, along with the image on the banner for that week. Contribution levels always increased with the transition from flowers to eyes, and decreased with the transition from eyes to flowers. A general linear model with factors image type (fixed) and week (covariate) fitted to log-transformed data explained 63.8% of the variance. There was a significant main effect of image type (eyes versus flowers: $F_{1,7}=11.551$, $p=0.011$) but not week ($F_{1,7}=0.074$, $p=0.794$). The interaction between image type and week was omitted from the model because it was not significant. On average, people paid 2.76 times as much in the weeks with eyes (mean $\pm$ s.e. $=0.417\pm0.081$ £ per litre) than with flowers (0.151 ± 0.030 £ per litre). There was no evidence that image type affected consumption.

[4] Bateson et al., 2006, Cues of being watched enhance cooperation in a real-world setting.

Step 3: Generate hypotheses

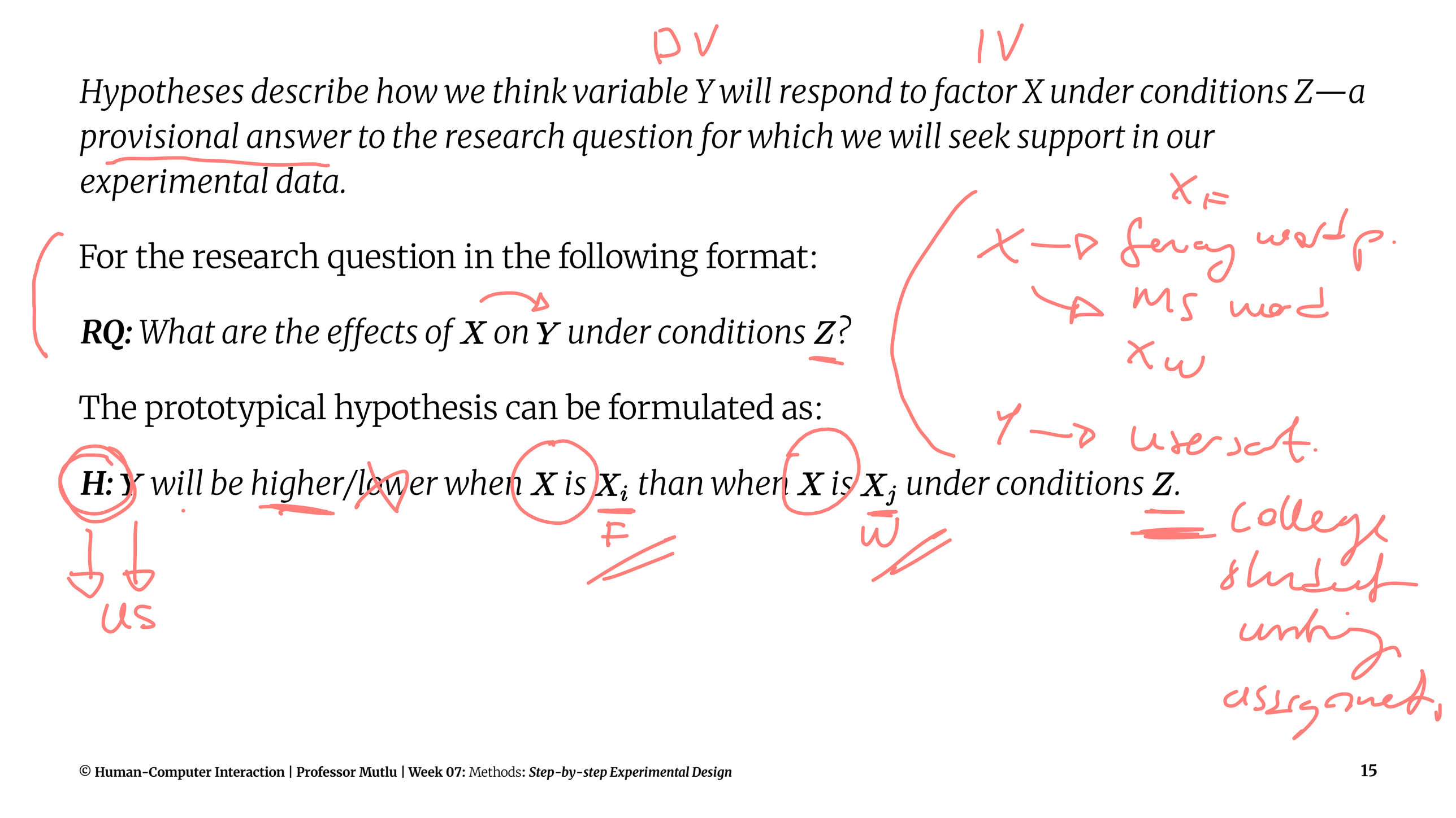
Hypotheses describe how we think variable Y will respond to factor X under conditions Z—a provisional answer to the research question for which we will seek support in our experimental data.

For the research question in the following format:

RQ: *What are the effects of* X *on* Y *under conditions* Z*?*

The prototypical hypothesis can be formulated as:

H: Y *will be higher/lower when* X *is* X_i *than when* X *is* X_j *under conditions* Z*.*

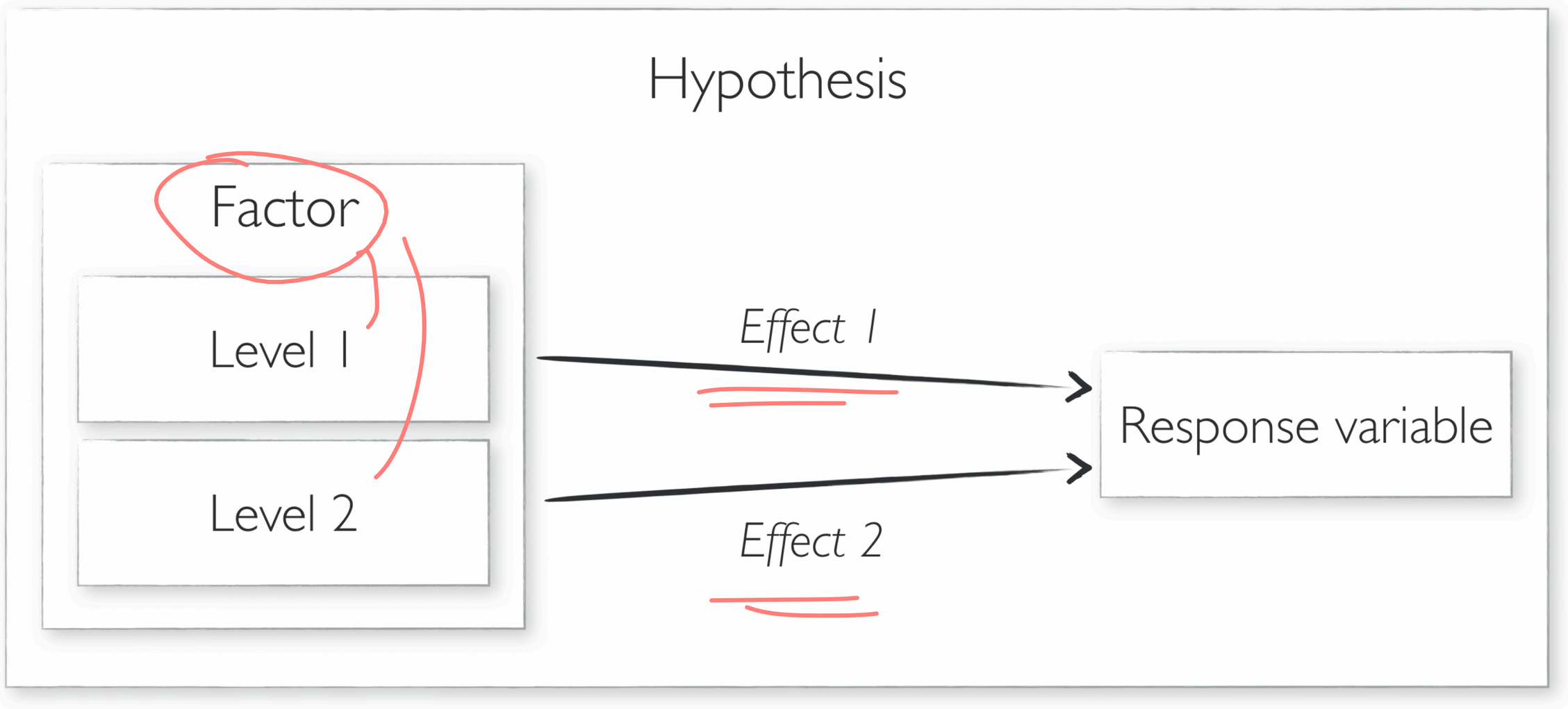


How do we come up with a provisional description of the relationship between X *and* Y*?*

Hypotheses can come from three sources:

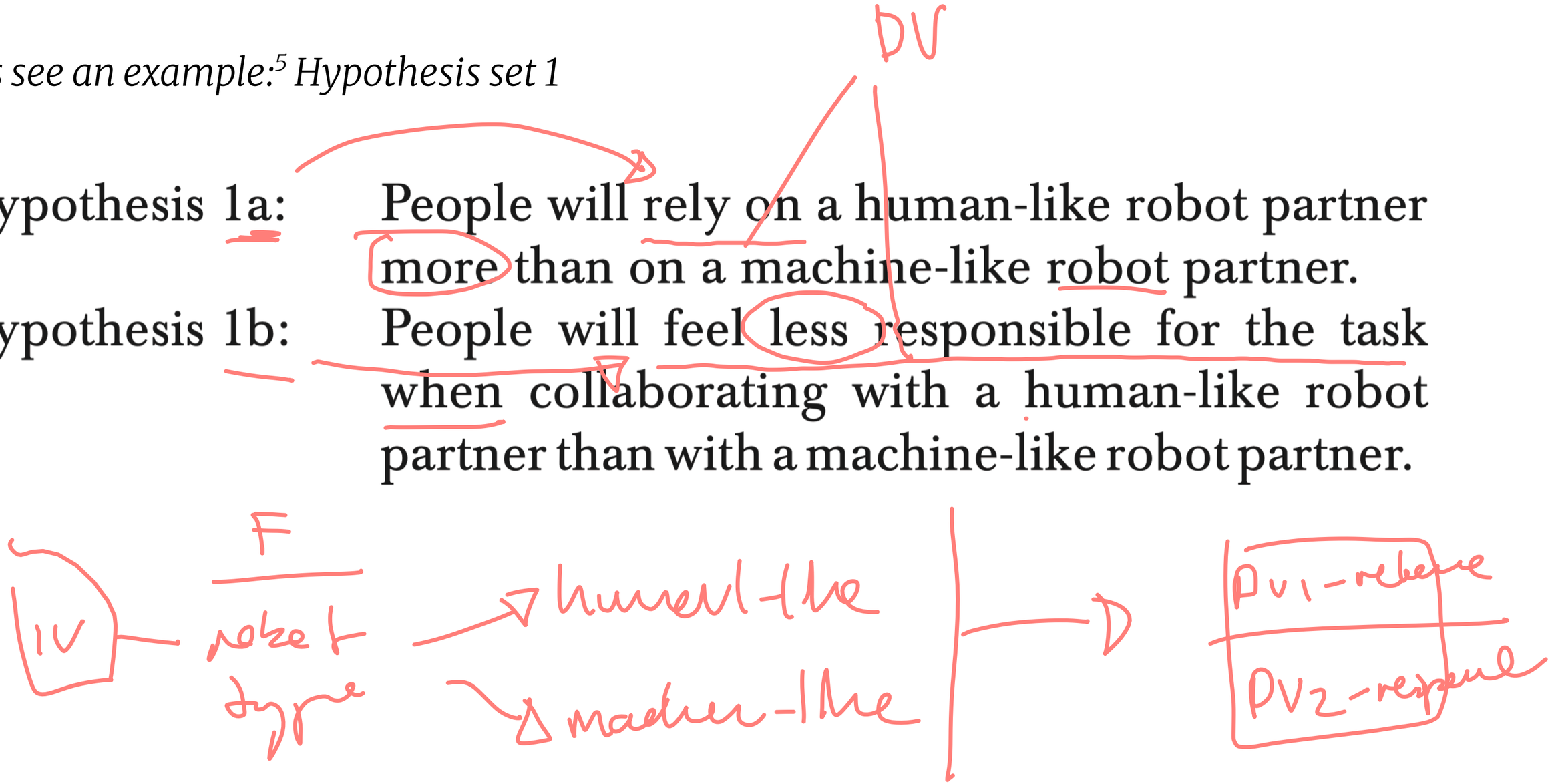
1. Results from exploratory studies
2. Existing theory in a different but related area
3. Logical reasoning with face validity

In all cases, hypotheses must be justified.

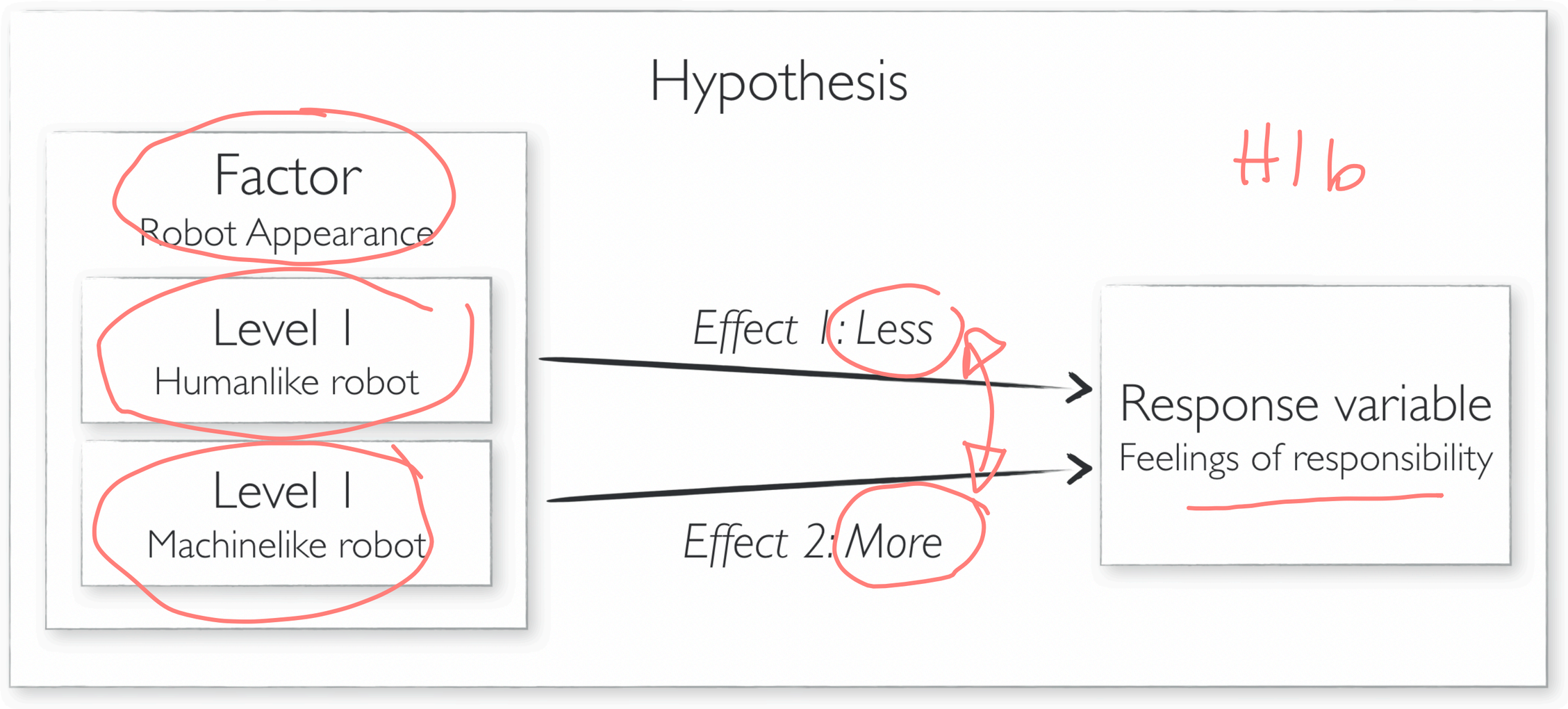

Hypothesis
Factor
Level 1
Level 2
Effect 1
Effect 2
Response variable

Let's see an example:[5] Hypothesis set 1

- Hypothesis 1a: People will rely on a human-like robot partner more than on a machine-like robot partner.
- Hypothesis 1b: People will feel less responsible for the task when collaborating with a human-like robot partner than with a machine-like robot partner.



[5] Hinds, 2004, Whose job is it anyway? A study of human-robot interaction in a collaborative task


Hypothesis
Factor
Robot Appearance
Level 1
Humanlike robot
Level 1
Machinelike robot
Effect 1: Less
Effect 2: More
Response variable
Feelings of responsibility
H1b

Let's see an example:[5] *Hypothesis set 2*

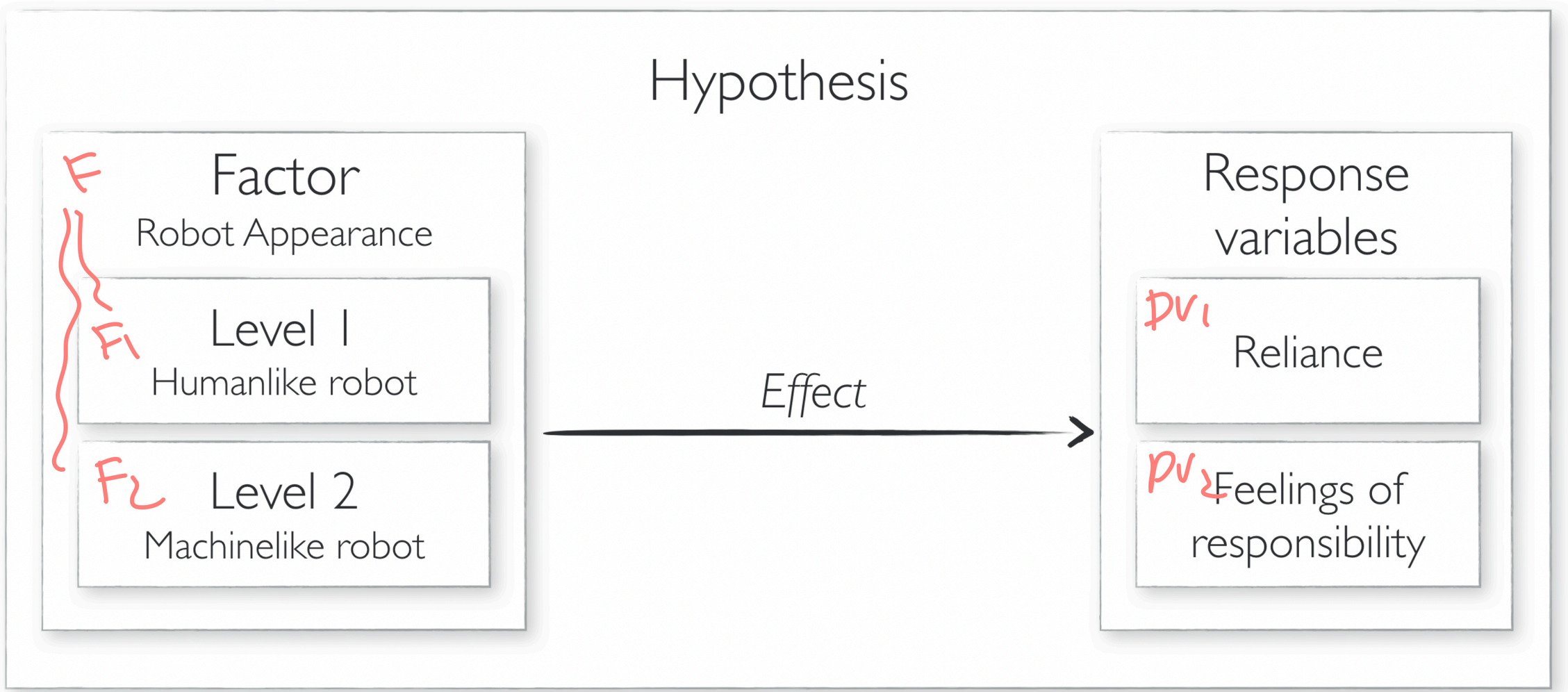
Human-like versus machine-like robots

Hypothesis 1a: People will rely on a human-like robot partner more than on a machine-like robot partner.

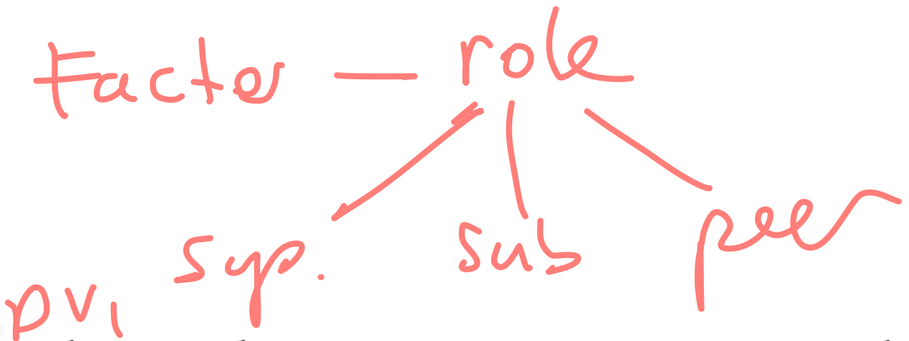
Hypothesis 1b: People will feel less responsible for the task when collaborating with a human-like robot partner than with a machine-like robot partner.

DVs

[5] Hinds, 2004, Whose job is it anyway? A study of human-robot interaction in a collaborative task


Hypothesis
Factor
Robot Appearance
Level 1
Humanlike robot
Level 2
Machinelike robot
Effect
Response variables
Reliance
Feelings of responsibility

Let's see an example:[5] Hypothesis set 3



Relative status of robot coworkers

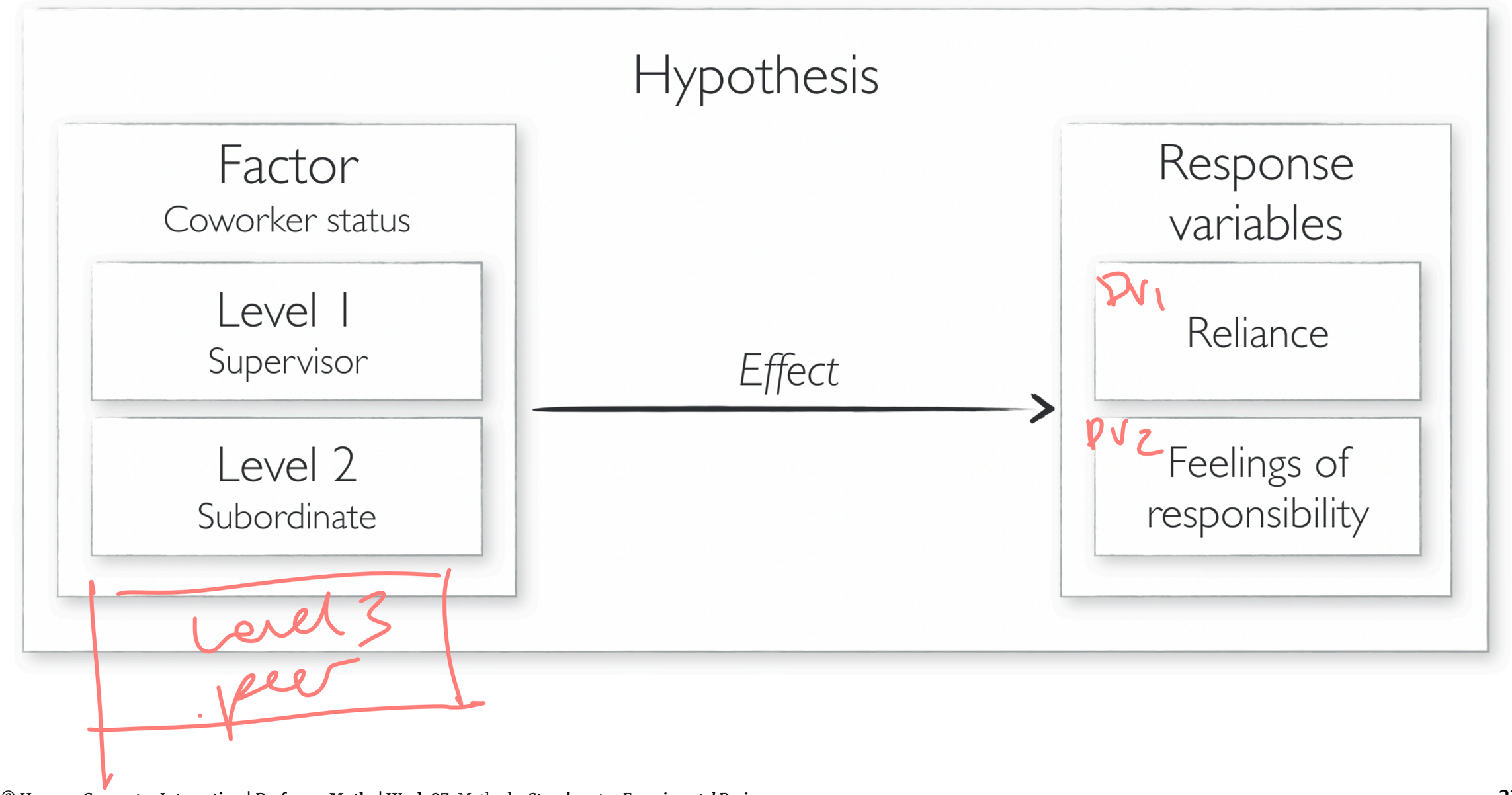
DV1

Hypothesis 2a: People will rely on the robot partner more when it is characterized as a supervisor than when it is characterized as a subordinate or peer.

DV2

Hypothesis 2b: People will feel less responsible for the task when collaborating with a robot partner who is a supervisor than with a robot partner who is a subordinate or peer.

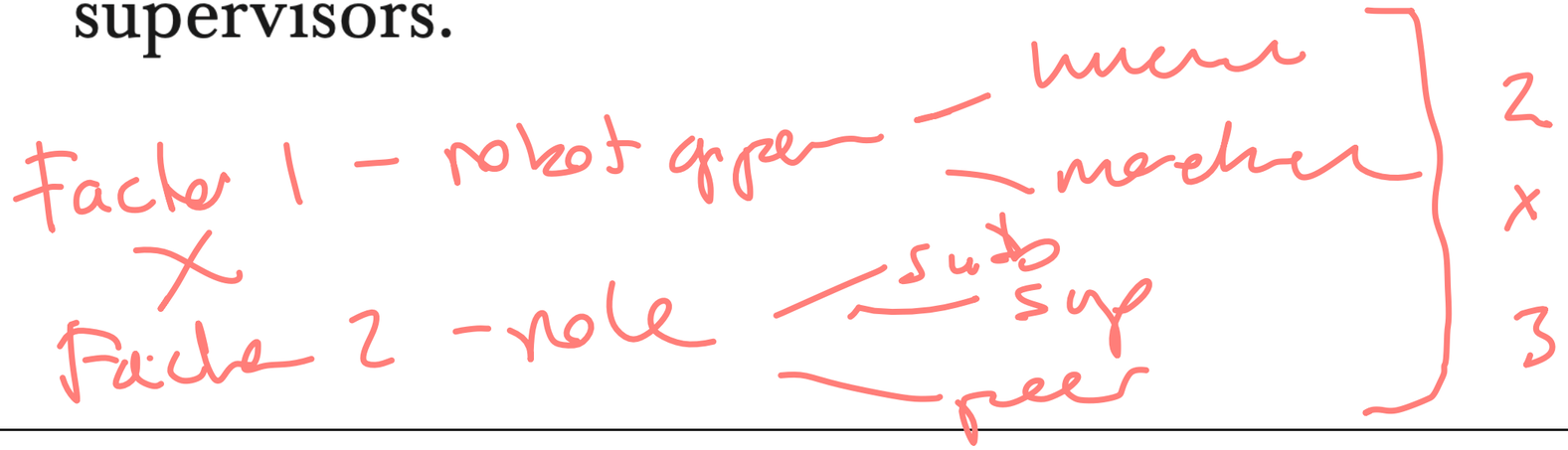
[5] Hinds, 2004, Whose job is it anyway? A study of human-robot interaction in a collaborative task


Hypothesis
Factor
Coworker status
Level 1
Supervisor
Level 2
Subordinate
Effect
Response variables
DV1
Reliance
DV2
Feelings of responsibility
level 3
peer

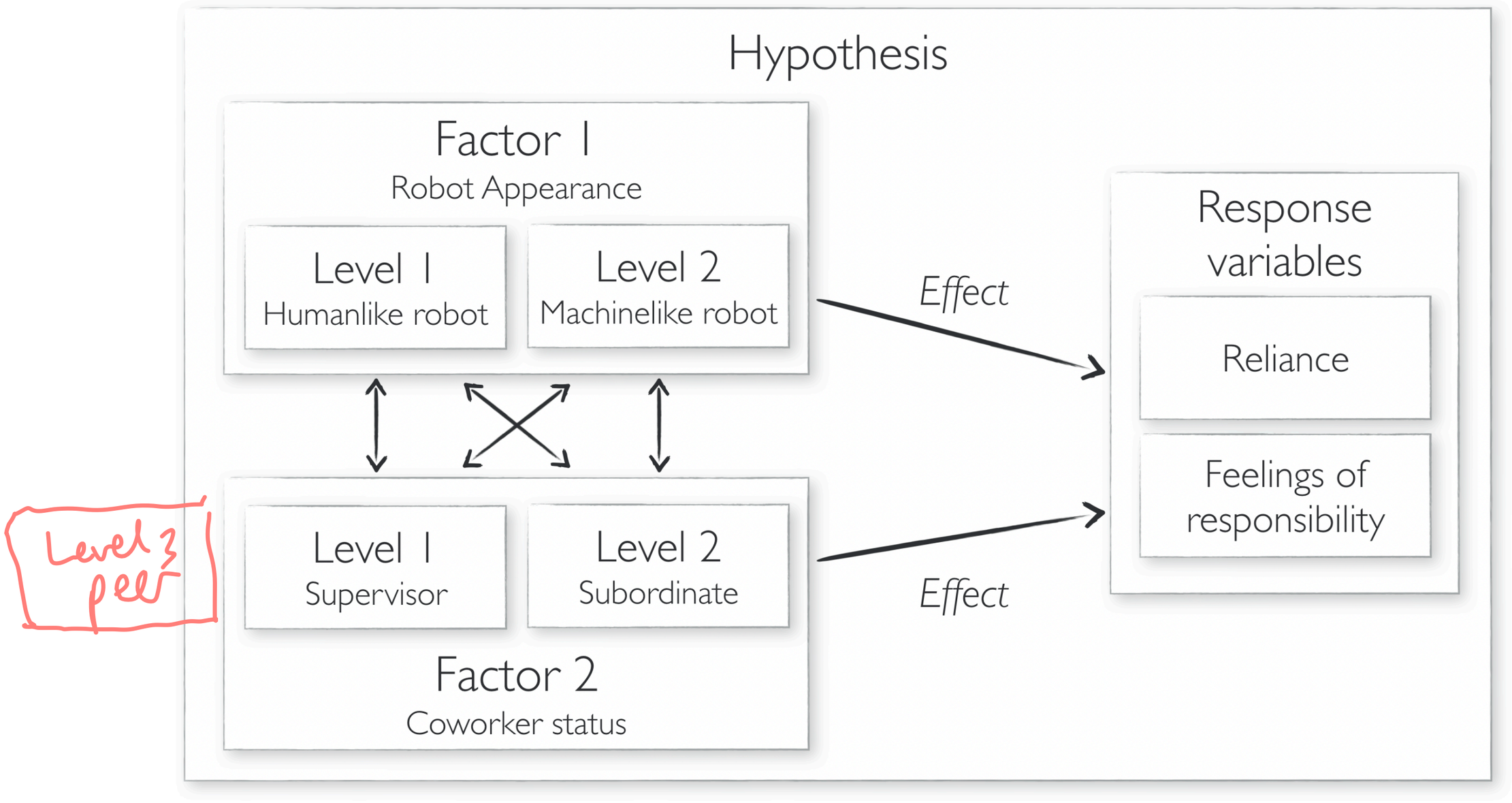
Let's see an example:[5] Hypothesis set 4

Interactions between human-likeness and status

Hypothesis 3: People will feel the greatest amount of responsibility when collaborating with machine-like robot subordinates as compared with machine-like robot peers and supervisors; and as compared with human-like robot subordinates, peers, and supervisors.

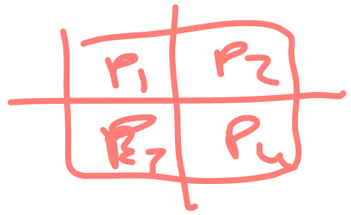


[5] Hinds, 2004, Whose job is it anyway? A study of human-robot interaction in a collaborative task


Hypothesis
Factor 1
Robot Appearance
Level 1
Humanlike robot
Level 2
Machinelike robot
Effect
Response variables
Reliance
Feelings of responsibility
Level 3 peer
Level 1
Supervisor
Level 2
Subordinate
Effect
Factor 2
Coworker status

Step 4: Determine experimental design

What is experimental design?



Definition: Experimental design refers to how participants are allocated to the different conditions.[6]

» **Simple designs** that vary one factor at a time are statistically inefficient and lead to wrong conclusions of factors interact

» **Factorial designs** that look at all combinations can simultaneously looks for effects of all factors but need more resources

» In general, factorial designs are recommended; 2^k designs are best

3×3
2×2×2 ✓

5 → 2×2×2×2×2 → 32
16

[6] Simply Psychology

What are our options?[7]

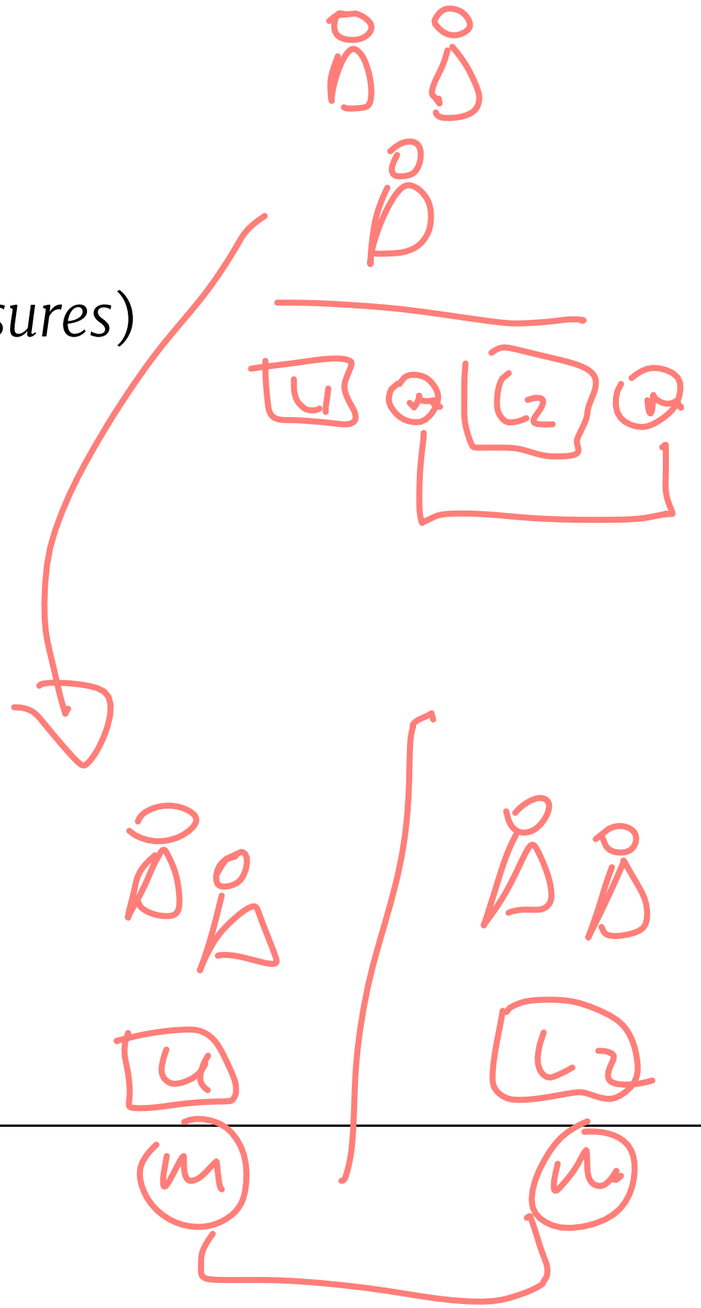
» **Within-participants** (also called *repeated measures*)

» **Between-participants** (also called *independent measures*)

» **Mixed-model** (also called *split-plot*)

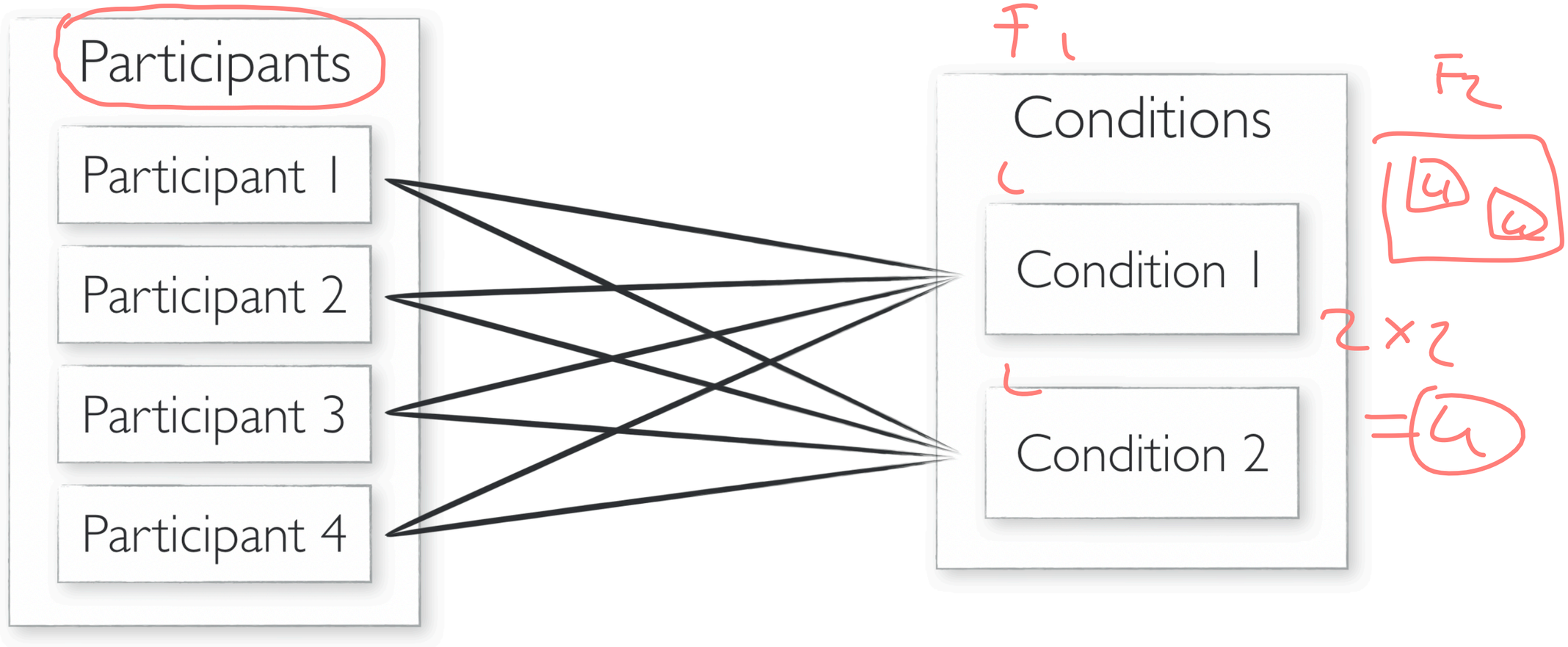
F1 → between

F2 → within



[7] There are other alternatives, e.g., *matched pairs*, but we will not cover them in this class.

*Recap: in **within-participants** design, all participants observe all levels of the manipulated factor.*

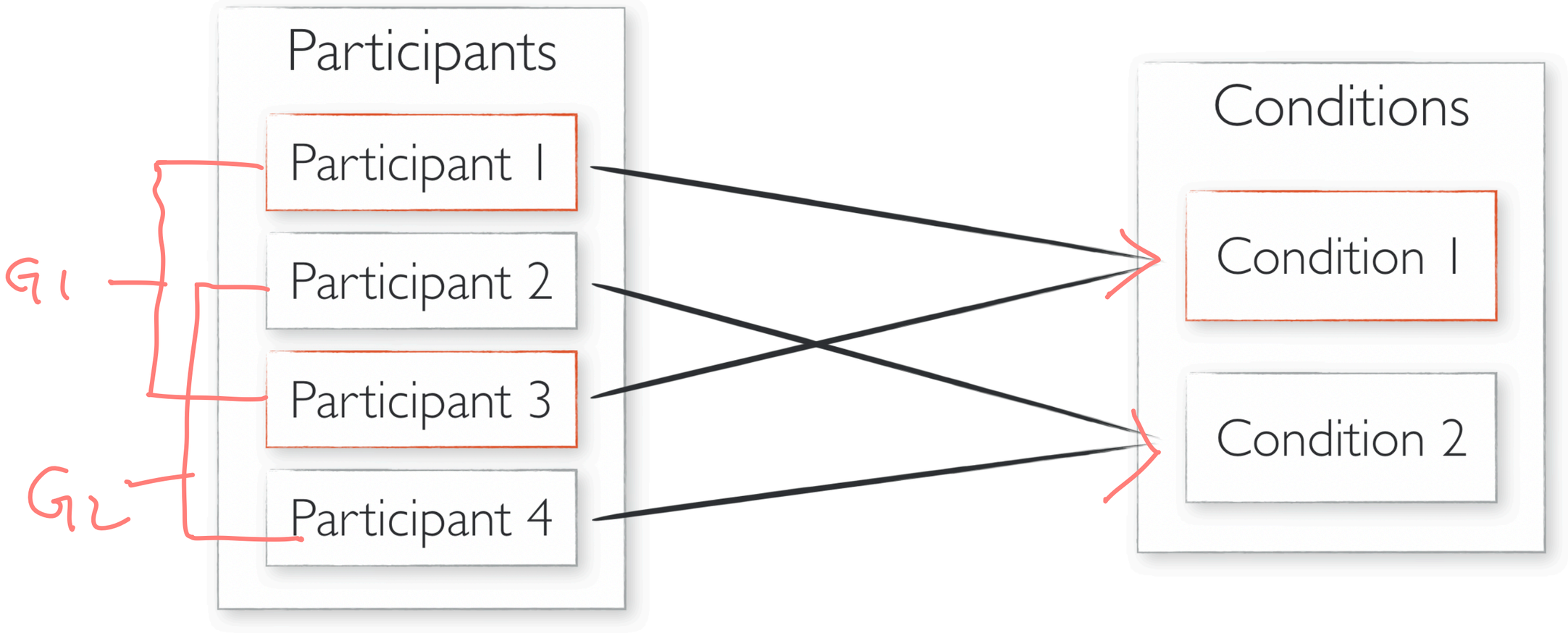


Example within-participants-design experiment:[8]

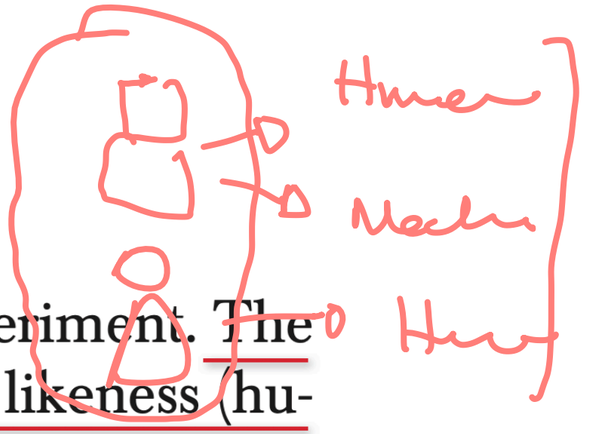
Study 2 was a 2 (condition: e-mail vs. voice) × 2 (accuracy: anticipated vs. actual) fully within-group factorial, with the dyad as the level of analysis. Because participants communicated different numbers of sarcastic statements, perceived and actual accuracy were converted to a percentage. Responses from one group were over 3 *SD*s away from the mean on several dependent variables and were excluded from the analysis, yielding a final sample size of 29 dyads.

[8] Kruger, 2005, Egocentrism over e-mail: Can we communicate as well as we think?

*Recap: in **between-participants** design, participants are divided into subgroups, and each subgroup observes one level of the manipulated factor.*



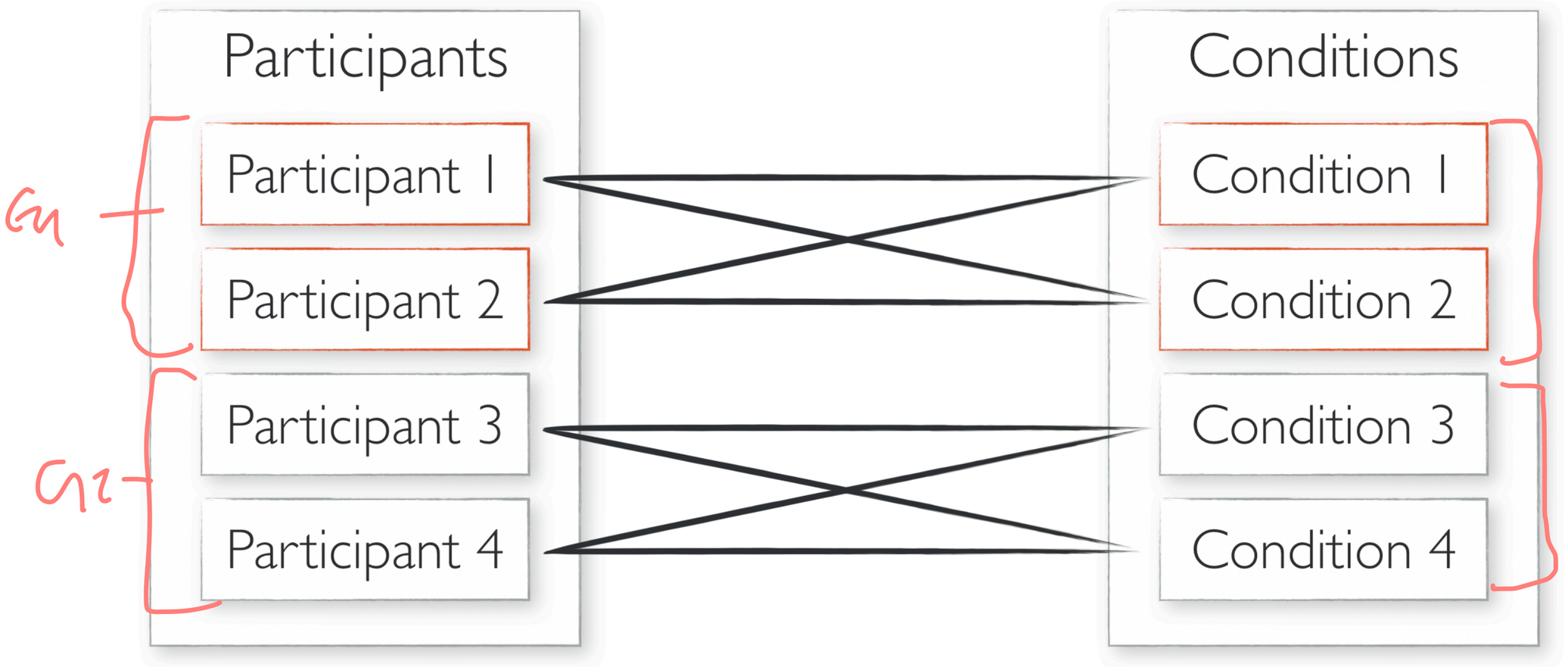
Example between-participants-design experiment:[5]



To test our hypotheses, we conducted a 3 ×3 laboratory experiment. The experiment was a between-subject design, manipulating human likeness (human, human-like robot, machine-like robot) and status (subordinate, peer, supervisor) with the human condition as the baseline. Each participant was asked to collaborate on a task with a confederate who reflected one of the nine cells in the design. The confederate used the same script for all conditions and was unaware of the status manipulation. In the robot conditions, we used a *Wizard of Oz* approach in which the robot was teleoperated, appearing to be operating autonomously. The same man teleoperated and spoke for the robot in the two robot conditions, and he acted as the human confederate. The experiment was videotaped with cameras suspended from the ceiling of the experimental lab.

[5] Hinds, 2004, Whose job is it anyway? A study of human-robot interaction in a collaborative task

*Recap: in **mixed-model** design, some factors are treated as within-participants and some factors are treated as between-participants.*



Example mixed-model-design experiment:[8]

> Our primary prediction was that overconfidence would be greater when participants communicated over e-mail than when participants communicated with their voice. To test this prediction, we conducted a 2 (accuracy: anticipated vs. actual) × 2 (order: Round 1 vs. Round 2) × 2 (acquaintanceship: stranger vs. friend) × 3 (medium: e-mail vs. voice-only vs. face-to-face) mixed-model ANOVA with the dyad as the level of analysis. The first two factors in this design were within-participants variables, and the second two were between-participants variables.

[8] Kruger, 2005, Egocentrism over e-mail: Can we communicate as well as we think?

How do we choose from among these options?

Choose **within-participants** designs when:

- Moderate *transfer effects*[9] and *demand characteristics*[10] are expected
- There are too many conditions that makes the study unfeasible due to the large number of participants required
- Inter-participant variance is expected to be high (e.g., when primary measures are performance based)

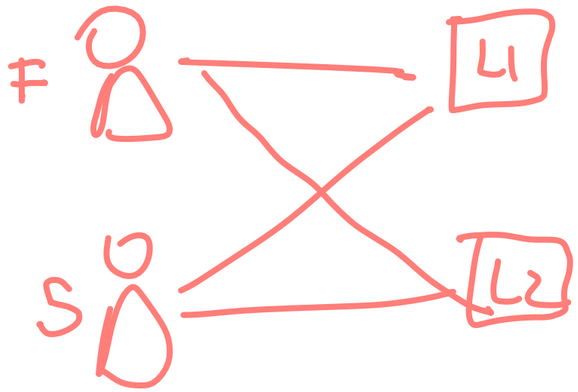
Provides more statistical power, needs fewer participants 😀; might impose bias due to these 👇 effects and can involve complex designs 😕

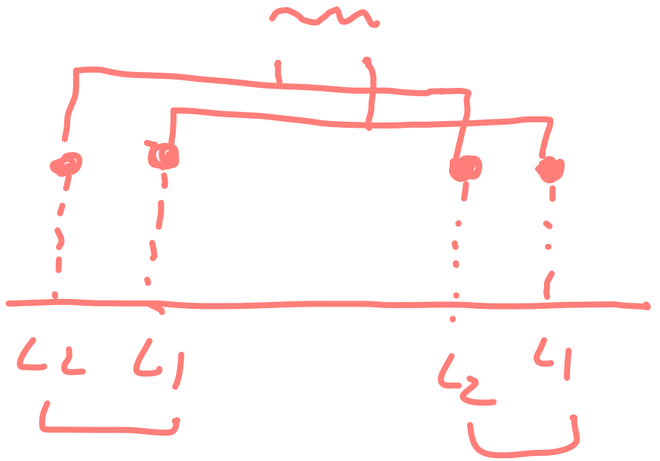
---

[9] **Transfer effects:** Taking part in earlier trials changes performance in the later trials due to learning, fatigue, etc.

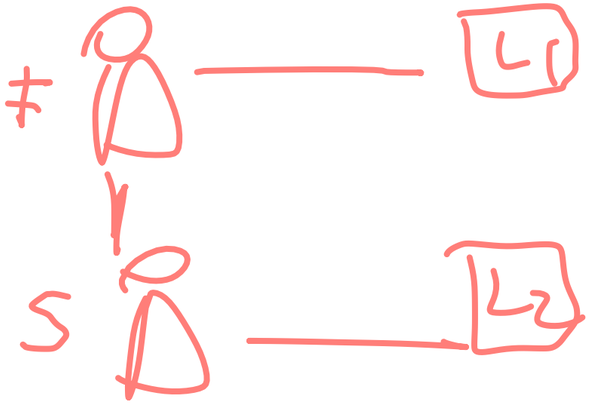
[10] **Demand characteristics:** Participants trying to question the purpose of the experiment.

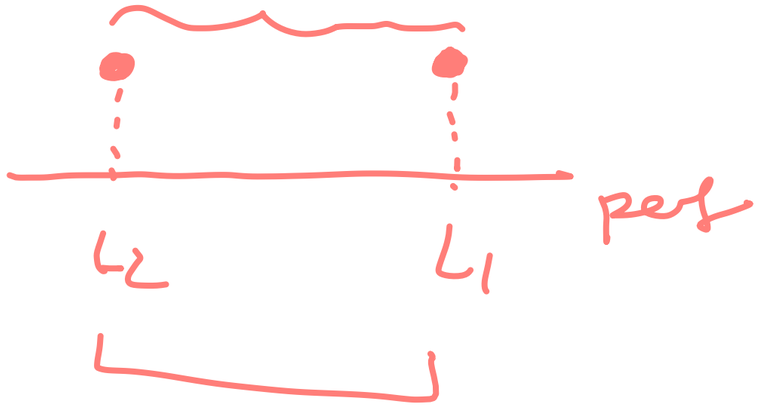
✓


L1
L2


L2
L1
L2
L1

✗


L1
L2


L2
L1
perf

Choose **between-participants** designs when:

- Severe *transfer effects* and *demand characteristics* are expected
- The required number of conditions and participants are feasible
- Inter-participant variance is expected to be moderate

Reduces bias by avoiding or alleviating undesirable experimental effects and easy to administer 😀; might result in high variance due to inter-participant variability 😕

Choose **mixed-model** designs when:

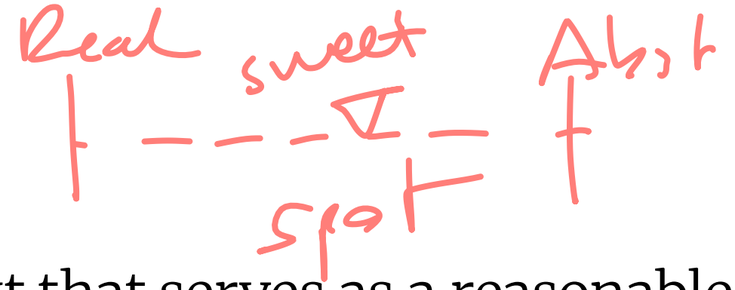
» Within-participants manipulation makes sense for some factors and between-participants manipulation makes sense for others

» A mixed design can be feasibly administered

Draws on the strengths of body designs 😀; can be difficult to administer, analyze, and interpret 😕

Step 5: Develop experimental task & procedure

What is an experimental task?



Definition: An experimental context that serves as a reasonable representation of real-world cognitive, social, and organizational situations that allows for generalizing to the real-world situation.

Experimental tasks:

- Must be a reasonable representation of the real-world context of interest (the **Z** defined in your research question)
- Must be relevant, reasonable, intutive, easy to interpret, and easy to control
- Must provide participants with appropriate motivational mechanisms to perform it as expected

Let's see an example:[11]

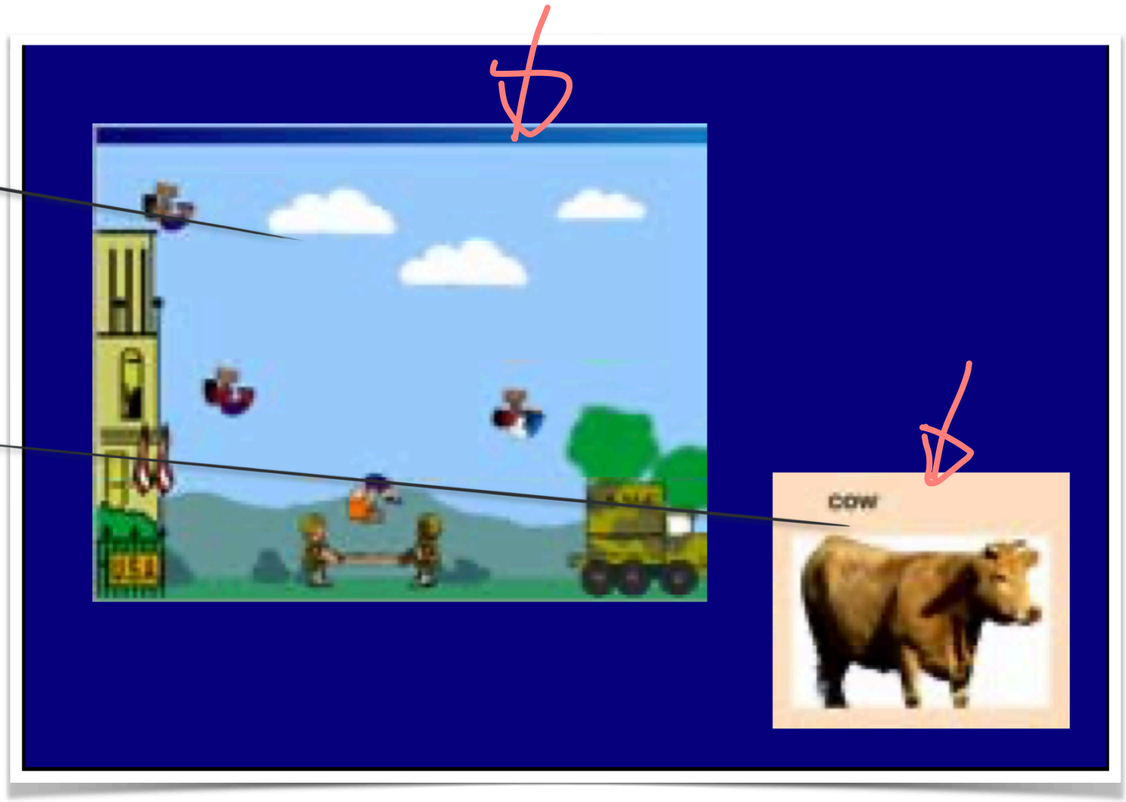
To investigate our hypotheses, we used the cooking tool selection questions exactly as they appeared in the pilot testing. The participants' goal was to select ten cooking tools needed to make a crème brûlée dessert. Participants selected the tool by clicking on the correct picture on a computer monitor. Each of the ten tools was displayed separately alongside five incorrect tools. The robot conversationally led the participant through the task, requesting each of the tools in turn, and answering participants' questions. Participants could ask the robot as many questions as they wished.

[11] Torrey et al., 2006, Effects of adaptive robot dialogue on information exchange and social relations.

Distractor tasks *are used to increase cognitive or perceptual demand on participants to understand how they respond to stimuli with limited resources.*[12]

Distractor task
(high attention demand)

Main task
(intermittently shown)



[12] Dabbish et al., 2005, Understanding email use: predicting action on a message.

Deceptive tasks *involve providing participants with a cover story that does not reflect experimental manipulations to minimize demand characteristics.*[13]

3.4 Procedure. One individual participated in each session. We instructed participants that they would be walking around a room and engaging in a memory test. They read the following paragraph:

> In the following experiment, you will be walking around in a series of virtual rooms. In the rooms with you will see a person. The person is wearing a white patch on the front of his shirt. His name is written on that patch. He is also wearing a similar patch on the back of his shirt. On the back patch, a number is written. Your job is to walk over to the person in the room and to read the name and number on his patches. First, read the back patch, and then read the front patch. Later on, we will be asking you questions about the names and numbers of the person in each room. We will also be asking you about their clothing, hair color, and eye color. When you have read the patches and examined the person in each room, we will ask you to step back to the starting point in the room. The starting point is marked by a piece of wood on the floor.

Our ostensible experimental task of reading and memorizing the agent's name and number motivated the participant to move within a relatively close range (1 m or less) of the agent so as to easily read the textual material. We felt that, by design, this secondary task would unwittingly cause the subject to move close enough to the avatar as to intrude potentially upon the hypothesized personal space bubble of this entity. Subsequently, the participant's movements would result from a competition between their desire to maintain an appropriate level of personal space and their need to accurately read the patches.

Debriefing

[13] Bailenson et al., 2001, Equilibrium theory revisited: Mutual gaze and personal space in virtual environments.

What is an experimental procedure?

Definition: An experimental procedure is a detailed description of the steps involved in administering the experiment to facilitate replicability.

The experimental procedure should include:

» Details of the task and the instructions participants received

» Participant's role in the task and the study

» The actions of the experimenter administering the study

» The research equipment used

» A timeline of when consent was obtained, measurements were taken, and compensation was provided

Let's see an example:[11]

When participants arrived at the experimental lab, the experimenter told the participant that the robot had been given "specific expertise" in cooking, and that "the robot will be talking to you about the tools needed to make a crème brûlée dessert."

The robot spoke aloud and also displayed its messages on a display on the robot's chest. The robot used Cepstral's Theta [18] for speech synthesis, and its lips moved as it spoke. The text also showed on the screen, as in Instant Messenger interfaces. The interface was identical to the interface in [26] except that the dialogue technology was improved further, as discussed in the next section of this paper.

[11] Torrey et al., 2006, Effects of adaptive robot dialogue on information exchange and social relations.

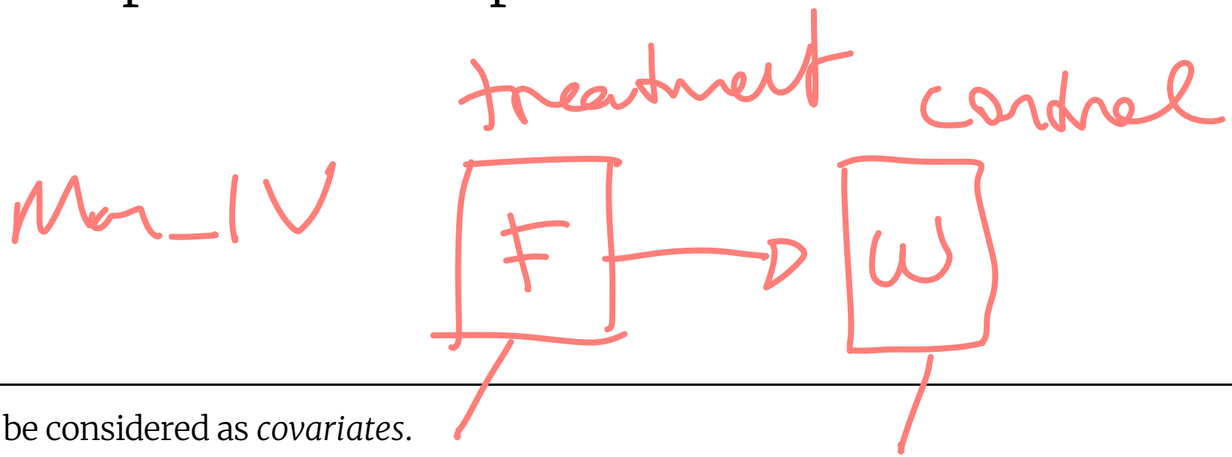
Step 6: Determine manipulations & measurements

What is being manipulated and what is being measured?

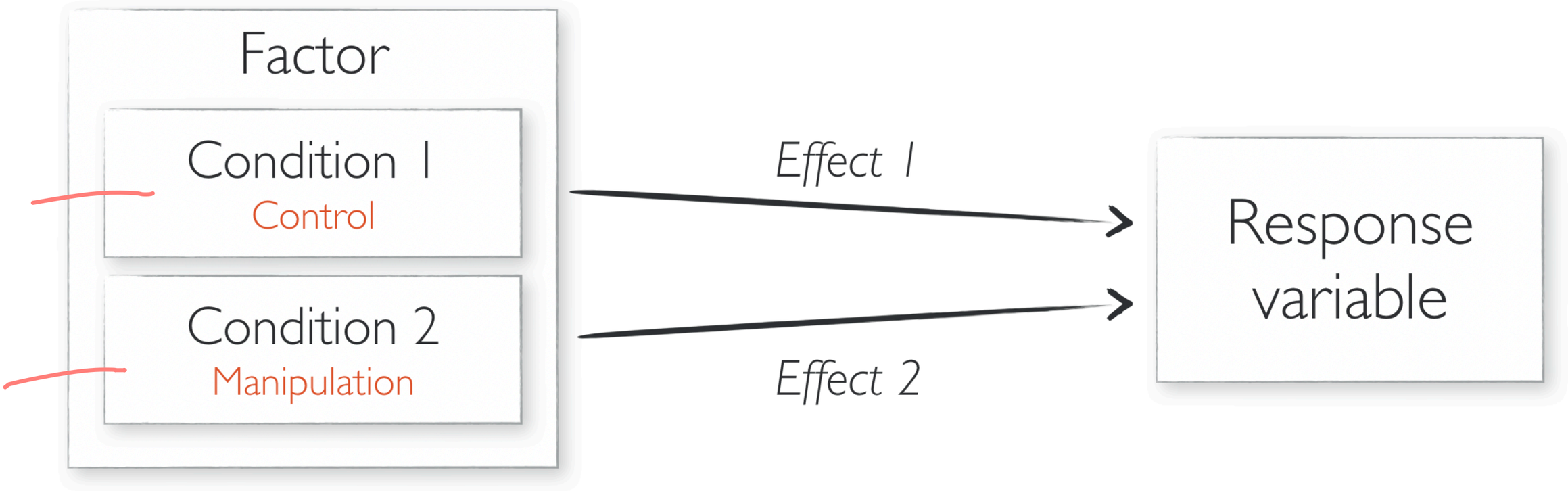
Independent variables can be *manipulated* (e.g., participants use interface 1 or interface 2) or *measured* (e.g., participants who are novices or experts). [14]

Manipulated independent variables usually involve *control* and *manipulation* (or *treatment*) levels.

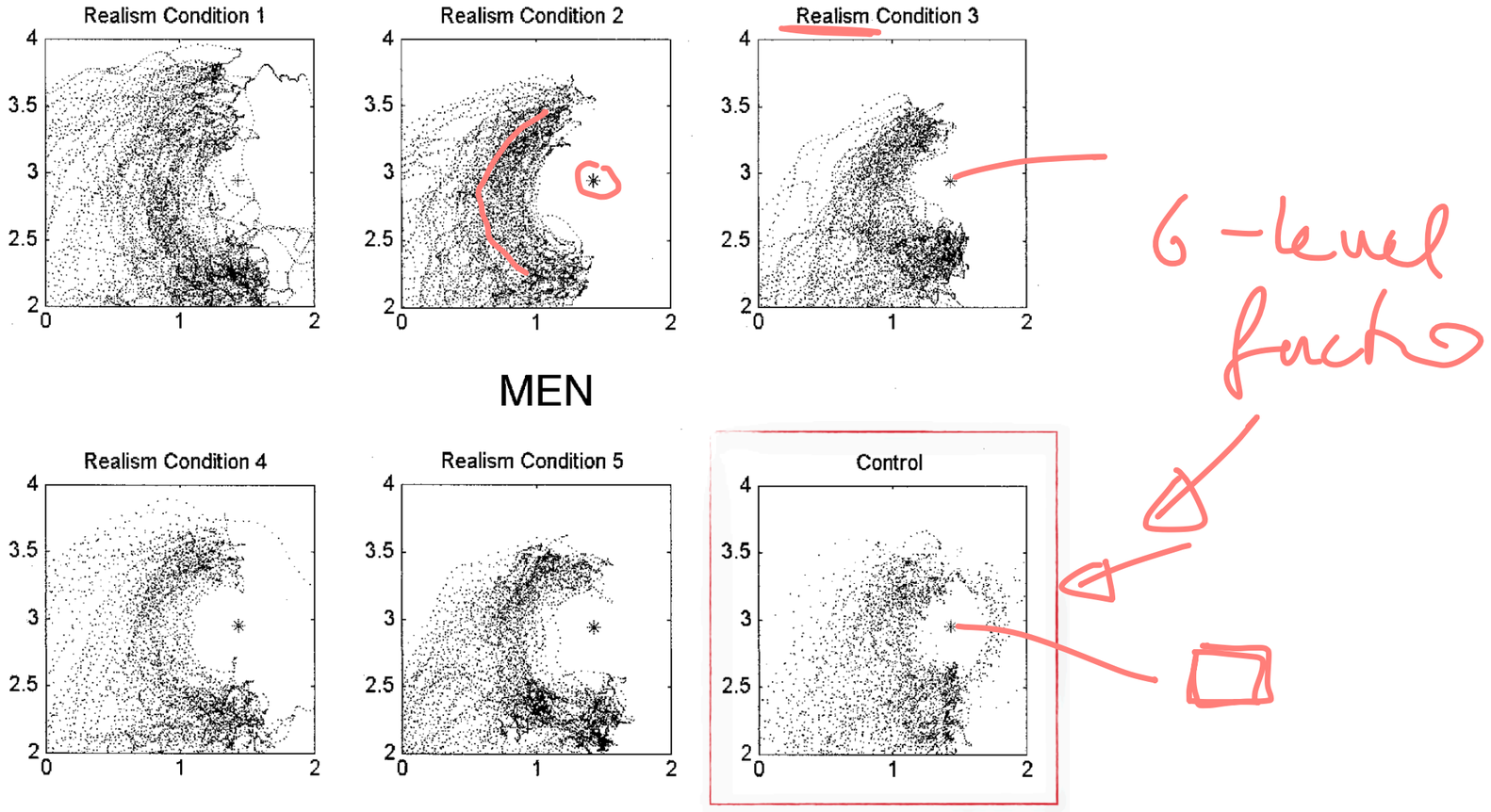
Control levels provide us with a *baseline* (lower bound) or a *gold-standard* (upper bound) against which to compare the manipulation.



[14] Measured independent variables can also be considered as *covariates*.


Factor
Condition 1
Control
Condition 2
Manipulation
Effect 1
Effect 2
Response variable

Let's see an example:[13]



[13] Bailenson et al., 2001, Equilibrium theory revisited: Mutual gaze and personal space in virtual environments.

How are independent or dependent variables measured?

Any variable, whether it is treated as a *factor* or a *response variable*, that is not explicitly manipulated must be measured.

Measures can capture participant performance, participant behavior, self-reported evaluations, physiological signals, and demographic characteristics.

More on these in the coming weeks.

Let's see an example:[15]

The following items were measured at the end of the fourth week. All measures are analytically distinctive and highly reliable (see Cronbach's α for each measure).

People's perception of AIBO as a developing creature was measured by two factors: (a) perceived development of AIBO and (b) perceived lifelikeness of AIBO. *Perceived development of AIBO* was measured based on the level of agreement (1 = *very strongly disagree*, 10 = *very strongly agree*) with the following statements: This AIBO has developed its skills over the course of four sessions because of my interaction with it; This AIBO's behavior has changed over the course of four sessions because of my interaction with it; This AIBO's intelligence has developed over the course of four sessions because of my interaction with it; This AIBO has matured over the course of four sessions because of my interaction with it; This AIBO has become more competent over the course of four sessions because of my interaction with it (α = .92). *Perceived lifelikeness of AIBO* was an index based on the level of agreement (1 = *describes very poorly*, 10 = *describes very well*) with the following adjectives describing AIBO: lifelike, machine-like (reverse coded), interactive, responsive (α = .76).

[15] Lee et al., 2005, Can a robot be perceived as a developing creature?

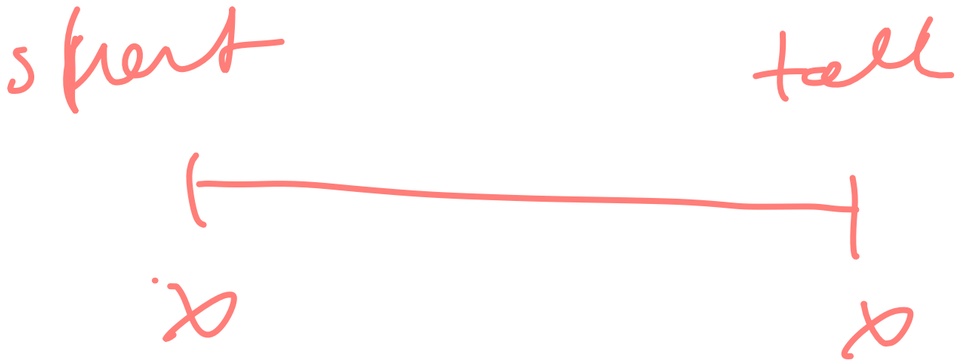
Step 7: Identify participants

How do we choose ***study participants****?*

Definition: A randomly sampled subpopulation of the general population that is relevant to the research question (expressed in the **Z**).

Study participants must be:

- Representative of the target population
- Sufficiently large to provide statistical power
- Balanced in measured factors



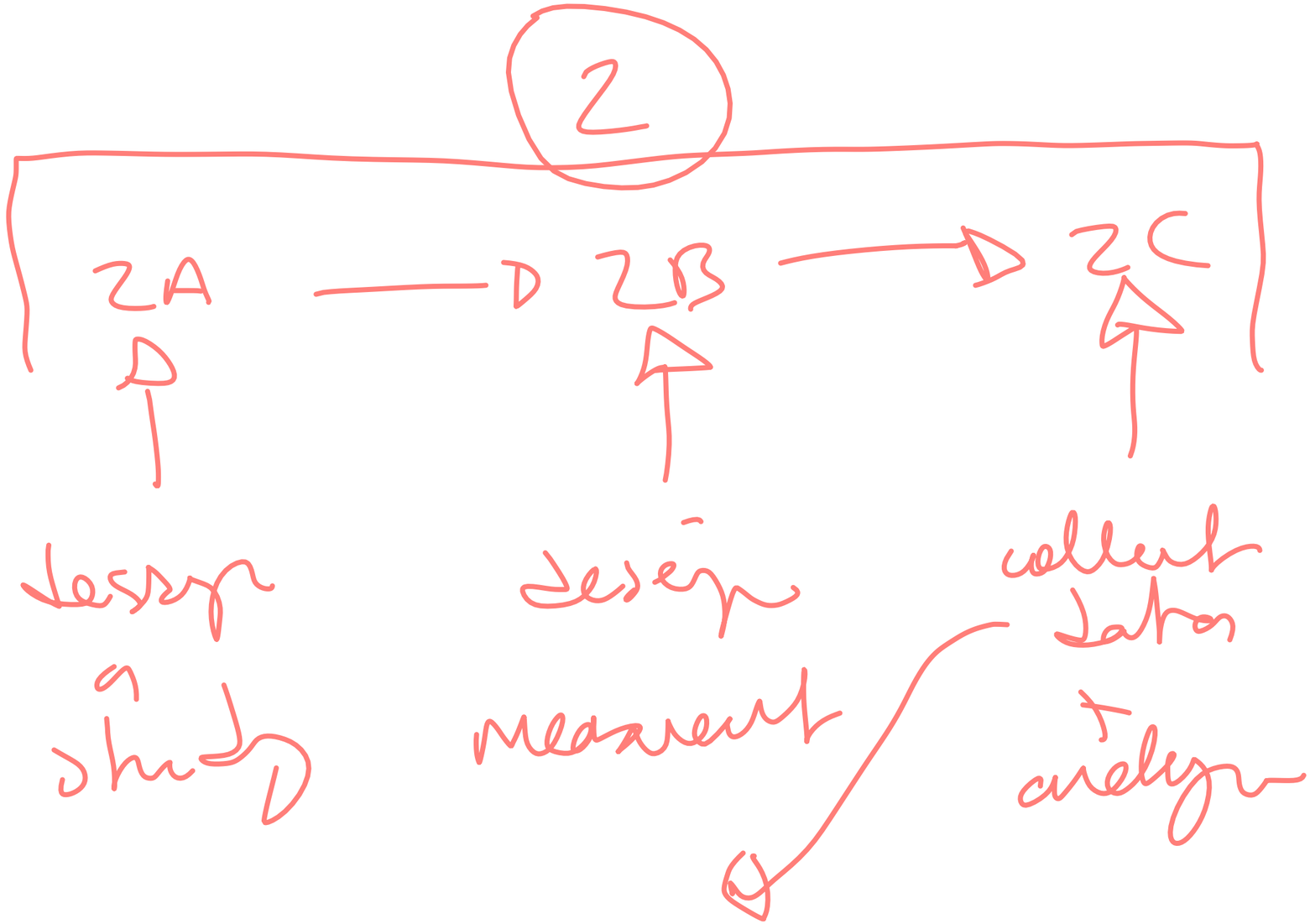
Let's see an example:[13]

2.3.3 Participants. Participants were recruited on campus and were either paid or given experimental credit in an introductory psychology class for participation. Four men and four women participated in each of the five gaze-behavior conditions, and six men and four women participated in the control condition, resulting in fifty total participants in the study. Participants' age ranged from 18 to 31.

[13] Bailenson et al., 2001, Equilibrium theory revisited: Mutual gaze and personal space in virtual environments.

Summary of the Steps

1. **Step 1:** Formulate research question
2. **Step 2:** Identify variables
3. **Step 3:** Generate hypotheses
4. **Step 4:** Determine experimental design
5. **Step 5:** Develop experimental task & procedure
6. **Step 6:** Determine manipulations & measurements
7. **Step 7:** Identify participants


2
2A
2B
2C
design
study
design
measurement
collection
data
analysis