University of Colorado
Boulder

Human-Robot Interaction

Step-By-Step Experimental Design

Professor **Dan Szafir**

Computer Science & ATLAS Institute
University of Colorado Boulder

Previously…

Fundamentals

Hypotheses
- Assumption, condition, and prediction

Independent vs dependent variable
- Which one is which?

Causality vs correlation

True- vs quasi-experimental design
- What is the difference? Why use one over the other?

Review

Between- vs within-participants design

- What is the difference? What are advantages / disadvantages of each?

Main vs interaction effect

- What are they? What do they imply about experimental design?

Counterbalancing, matching, and stratification

Single- vs double-blind studies

Why is this important?

DIS Review Paper 1:

R3 (not me):

"The paper seem to present the finding from the small-scale (4 families) study, **in a quantitative manner, rather than qualitatively, as would be expected from this type of study.** The paper presents few quotations from the participants, aiming rather to reduce their statements to generalizations, or to a line on a graph. Furthermore, the paper presents the difference among family members engagement with the robot, as distinct types, which seem to generalize types from a very small sample…"

Score: 1

R2 (me):

"The research appears to **suffer from poor methodological fit**. There has been a great deal of prior work investigating robot integration into human environments (including the appropriately-cited study by X). As a result, it is **unclear why we need another ethnographic study examining what is no longer a nascent research area**. Similarly, the results from the study appear to lack novelty…"

Score: 2

DIS Review Paper 2:

R3 (not me):

"The authors present an interesting study comparing participants' perception of X in 33 conditions, which demonstrated that people's perception of X is enhanced with Y. Given the **quantitative methods used in this study, I feel that more participants would provide stronger evidence**, or would like to have something in the paper justifying why there were **only five participants**."

Score: 2

R2 (me):

"The paper is **not well-motivated**. Why would such a [redacted] system be any more useful than a [commonly available object]? What research question is the manuscript trying to address? What is the primary contribution of the work?

…

No hypotheses are presented. The study recruited only 5 participants, leading to results **lacking in ecological validity.**"

Score: 1

Experimental Design Steps

Step I:
Formulate Research Question

Research Question

Should be specific enough

- Variables

Conditions under which the experiment will be performed should be understood

- Target population, experimental context, etc.

Should be informed by prior work

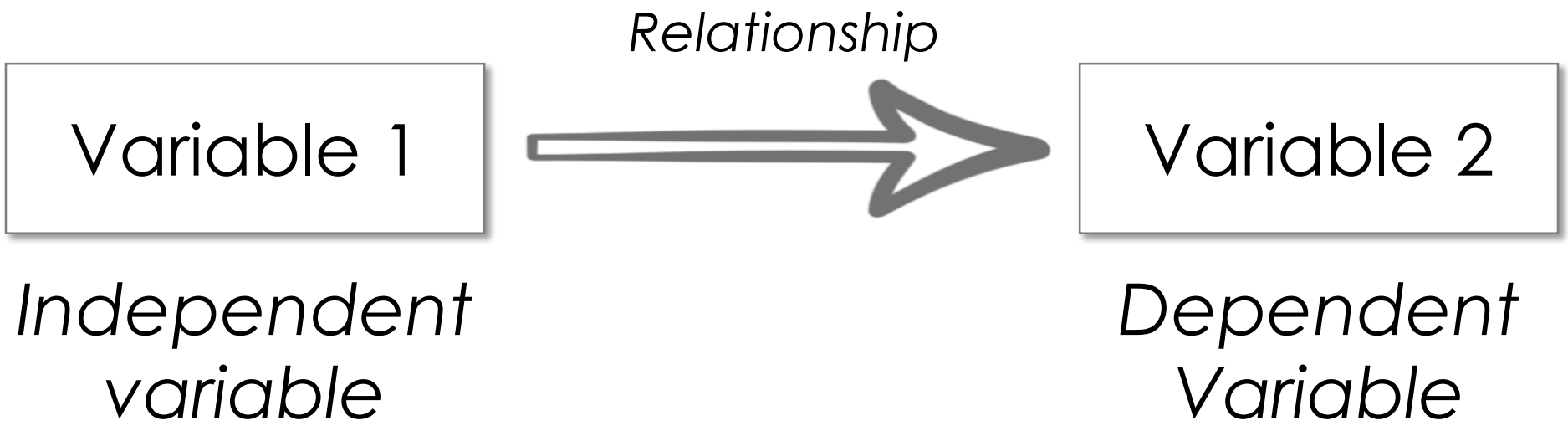
- Start thinking about methodological fit

Research Question

Questions?

Step II:
Identify Variables

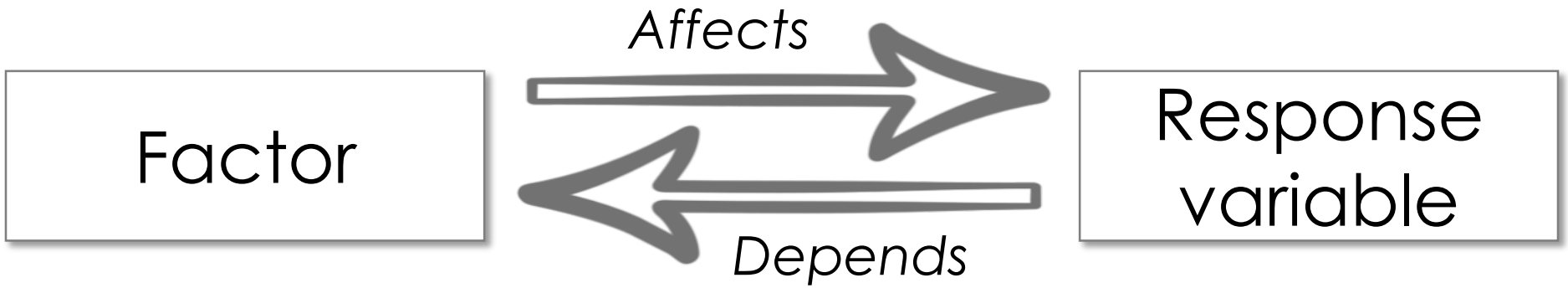
Variables


Relationship
Variable 1
Variable 2
Independent variable
Dependent Variable

Terminology

Factors

- Independent variables
- Factorial designs have at least two factors
  - E.g., 2 factors: gender + robot perception

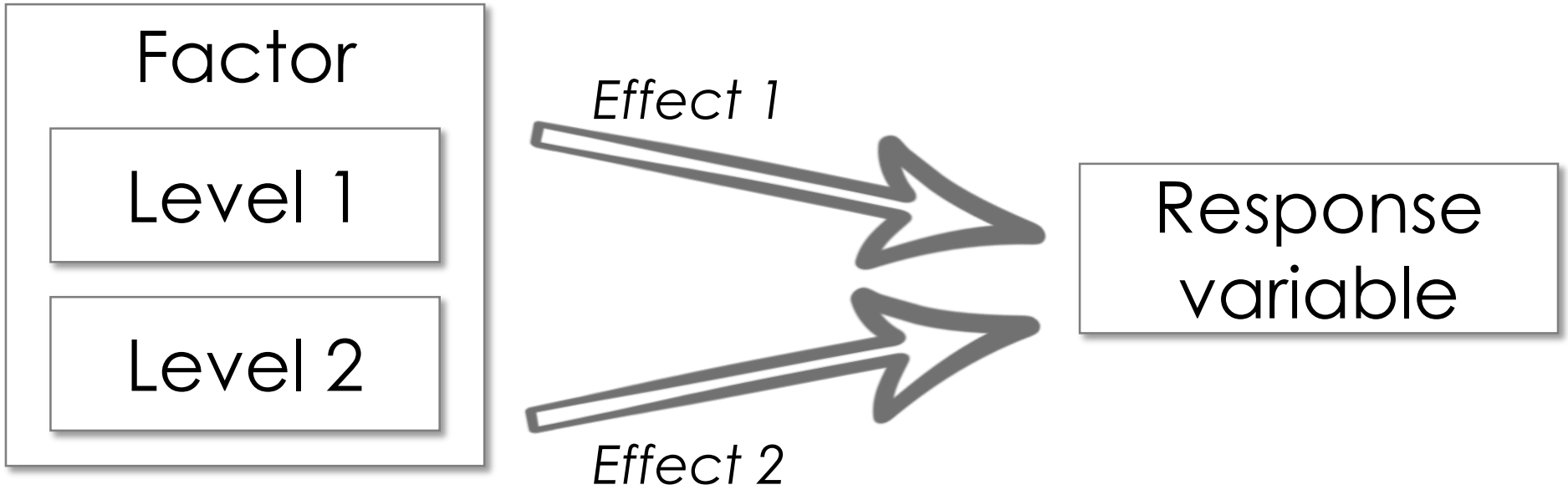


Terminology

Levels – also called "treatments"

Different "levels" a factor can take

E.g., gender: male/female; robot liking: low/medium/high



Factors vs Levels

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.

Nass C. and Lee. K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. Journal of Experimental Psychology: Applied, 7 (3): 171-181.

Factors vs Levels

2x2 design

2 factors, 2 levels

Factor 1: Computer voice personality

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

Fixed vs Random Factors

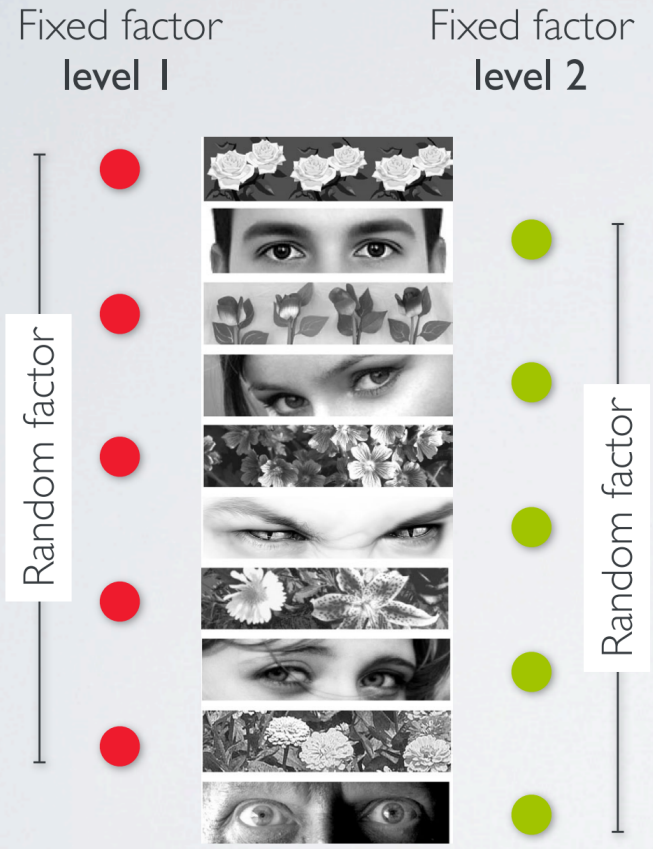
Fixed factors

- Levels under study are of interest

Random factors

- Levels under study are a random sample from a larger population
- The goal of the study is to generalize results to this larger population

Fixed vs Random Factors



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.

Bateson, M., Nettle, D., and Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. Biology Letters.

Questions?

Step III:
Generate Hypotheses

Hypotheses

Describes "how" you think variable **Y** will respond to factor **X** under conditions **Z**

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

X affects **Y** in **R way** under conditions **Z**

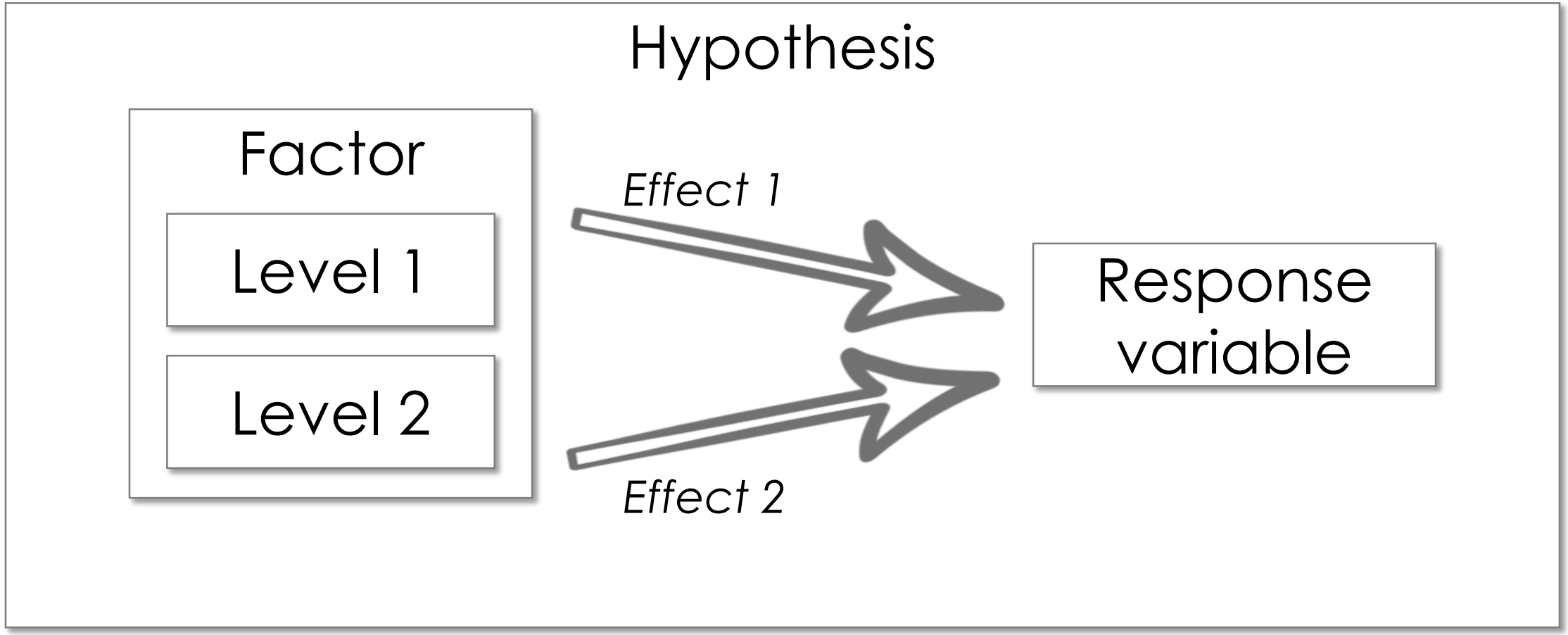
How is **R** determined?

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

Results from exploratory studies
Existing theory in a different but related area
Your intuition – *has to be justified*

X affects **Y** in **R way** under conditions **Z**

Generate Hypotheses



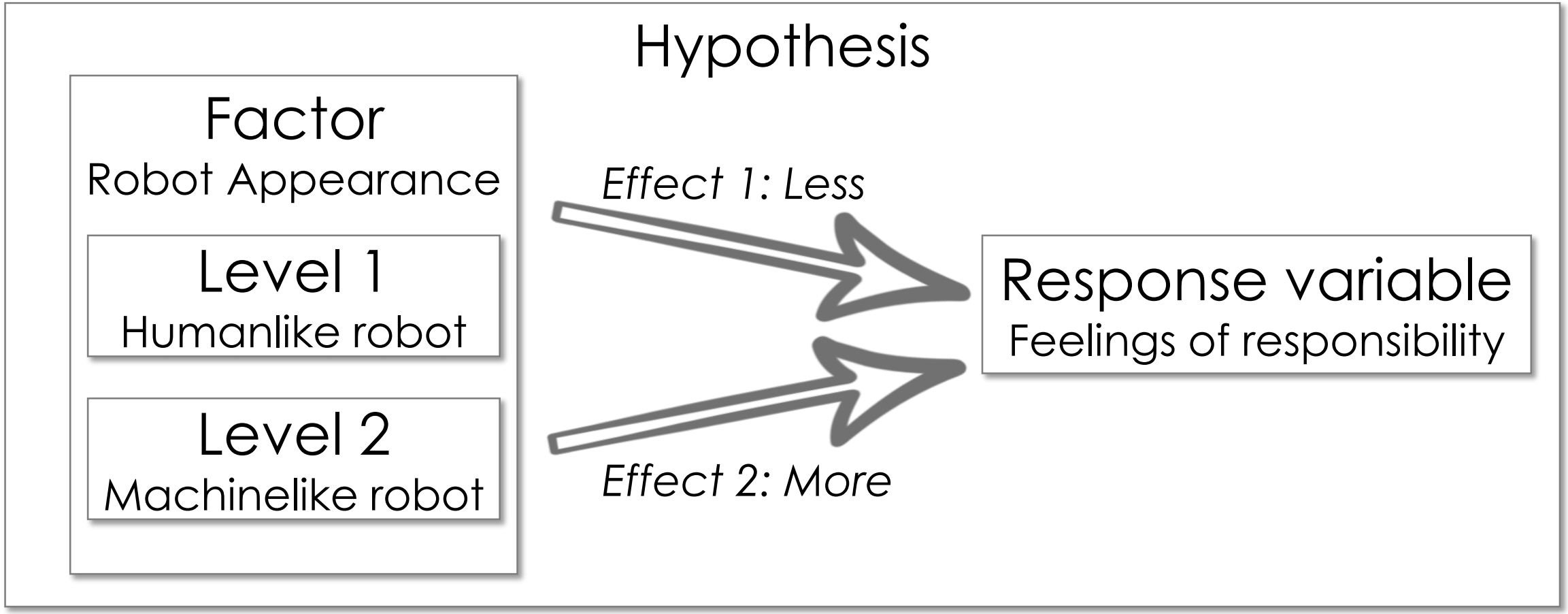
Generate Hypotheses

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.

Hinds, P.J., Roberts, T.L., and Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. Human-Computer Interaction, 19, 151-181.

Generate Hypotheses



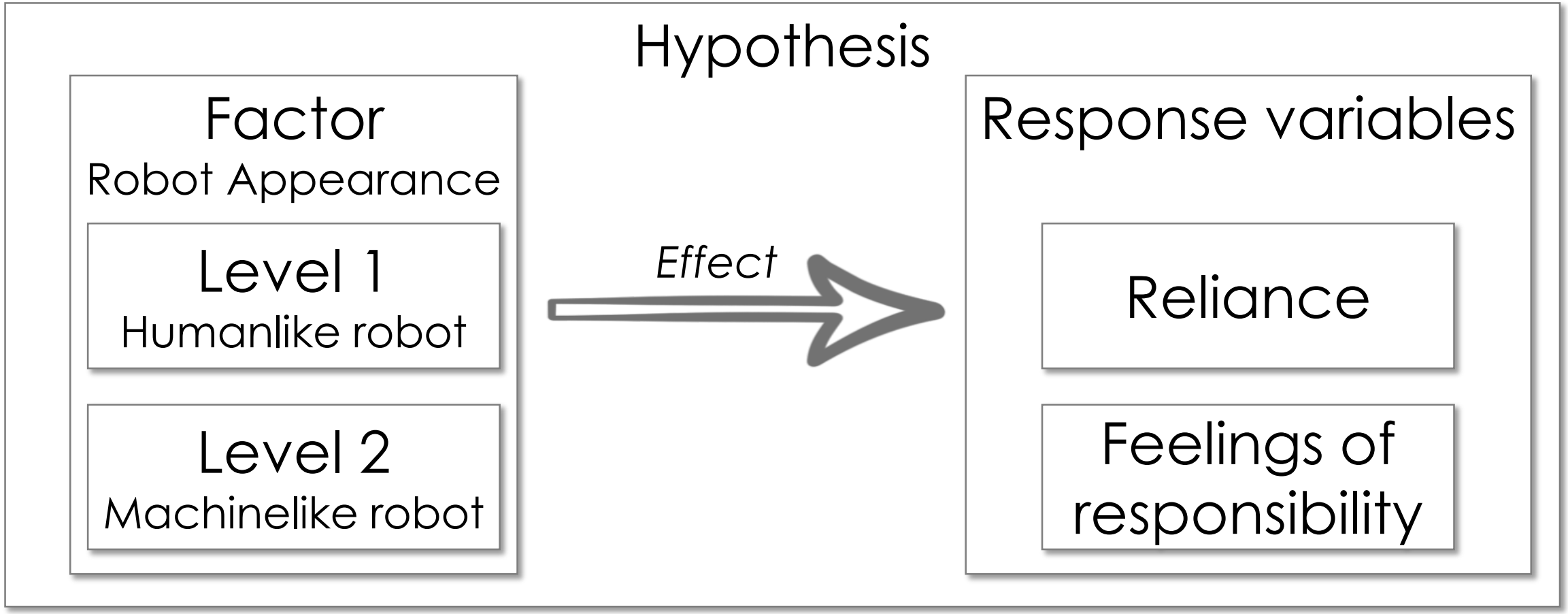
Generate Hypotheses

Human-like verses 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.

Generate Hypotheses



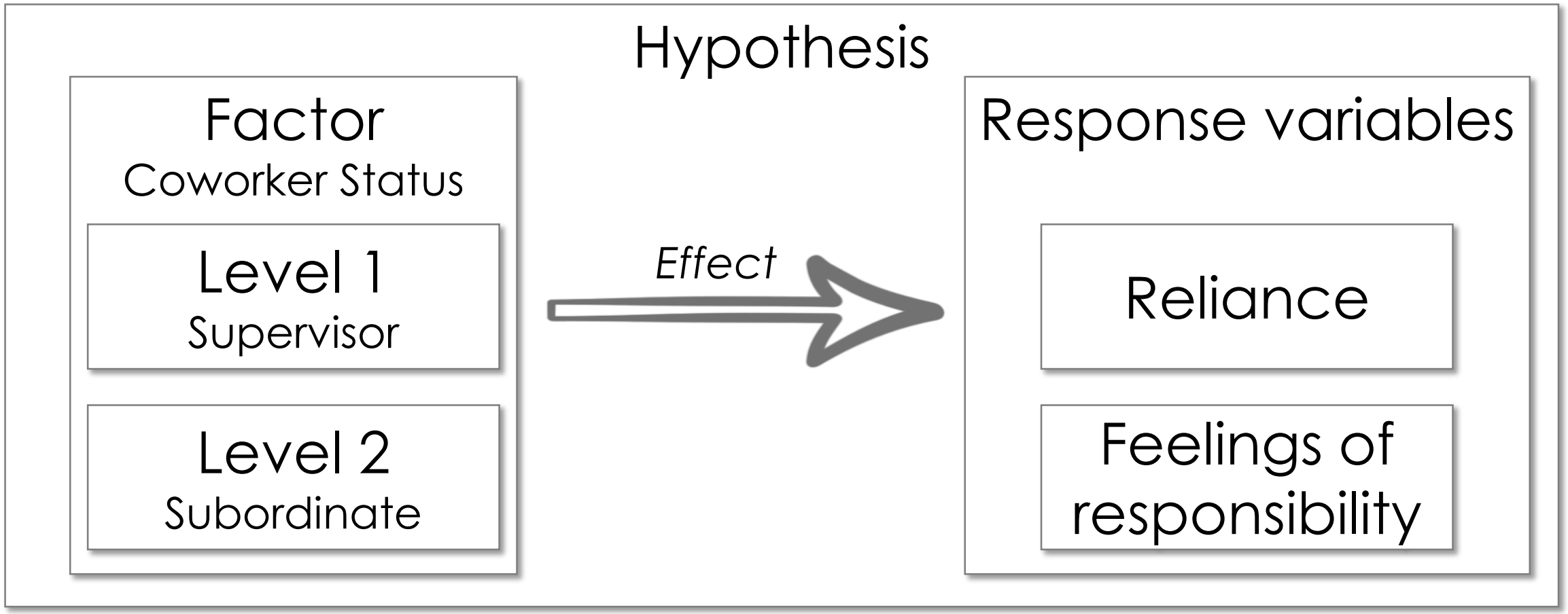
Generate Hypotheses

Relative status of robot coworkers

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.

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

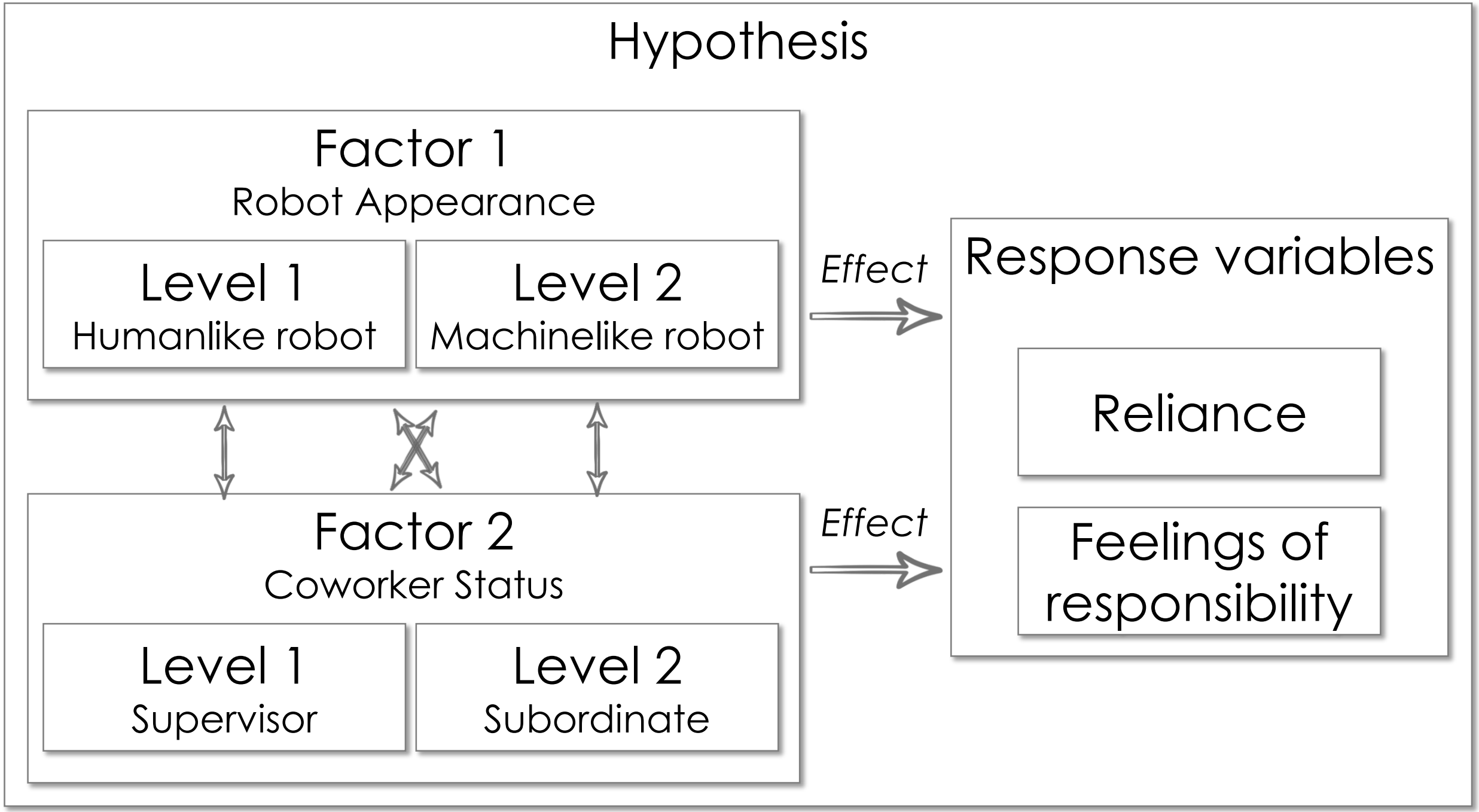
Generate Hypotheses



Factorial Designs

Interaction 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


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

Null Hypotheses

A statement that one seeks to nullify with evidence to the contrary

- Typically says the phenomenon being studied produces no effect

Example:

Hypothesis H_1: The presence of loud music will have a negative effect on student learning

Null Hypothesis H_0: Loud music will have no effect on student learning

Statistical Errors

Type 1 Error
- Incorrect rejection of the null hypothesis
- E.g., false positive

Type 2 Error
- Incorrectly failing to reject the null hypothesis
- E.g., false negative

Example:
- H_1: Evidence shows that person is guilty
- H_0: Person is innocent
- Type 1 error: convicting person if they really are innocent
- Type 2 error: letting person go free if they are actually guilty

Statistical Errors

To determine if you have made a Type I or II error, you must compare your decision with ground truth

- But we rarely, if ever, can know the ground truth!

We use statistical processes to balance our **risk** of making Type I and II errors

- e.g., selection of alpha level
- Will discuss more as we cover statistical inference

Questions?

Step IV: *Determine Experimental Design*

Experimental Complexity

Simple designs – vary one factor at a time

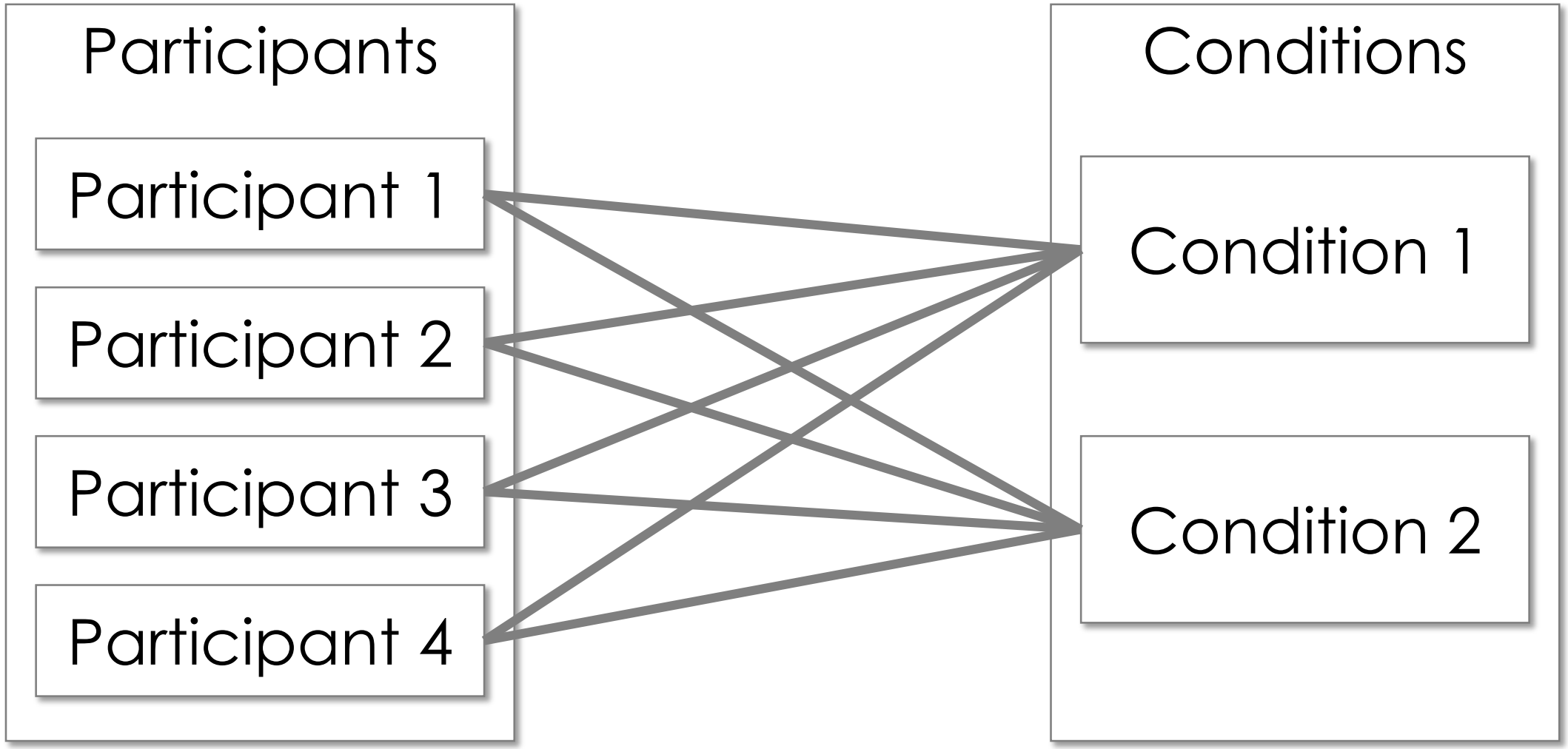
- Not statistically efficient (high possibility of Type 1 errors)
- Wrong conclusions if factors interact
- Overall, factorial designs are preferred

Experimental Complexity

Factorial designs – look at all combinations

- Can find the effect of all factors
- Needs too many resources
- 2^k is often most feasible

Within-participants design

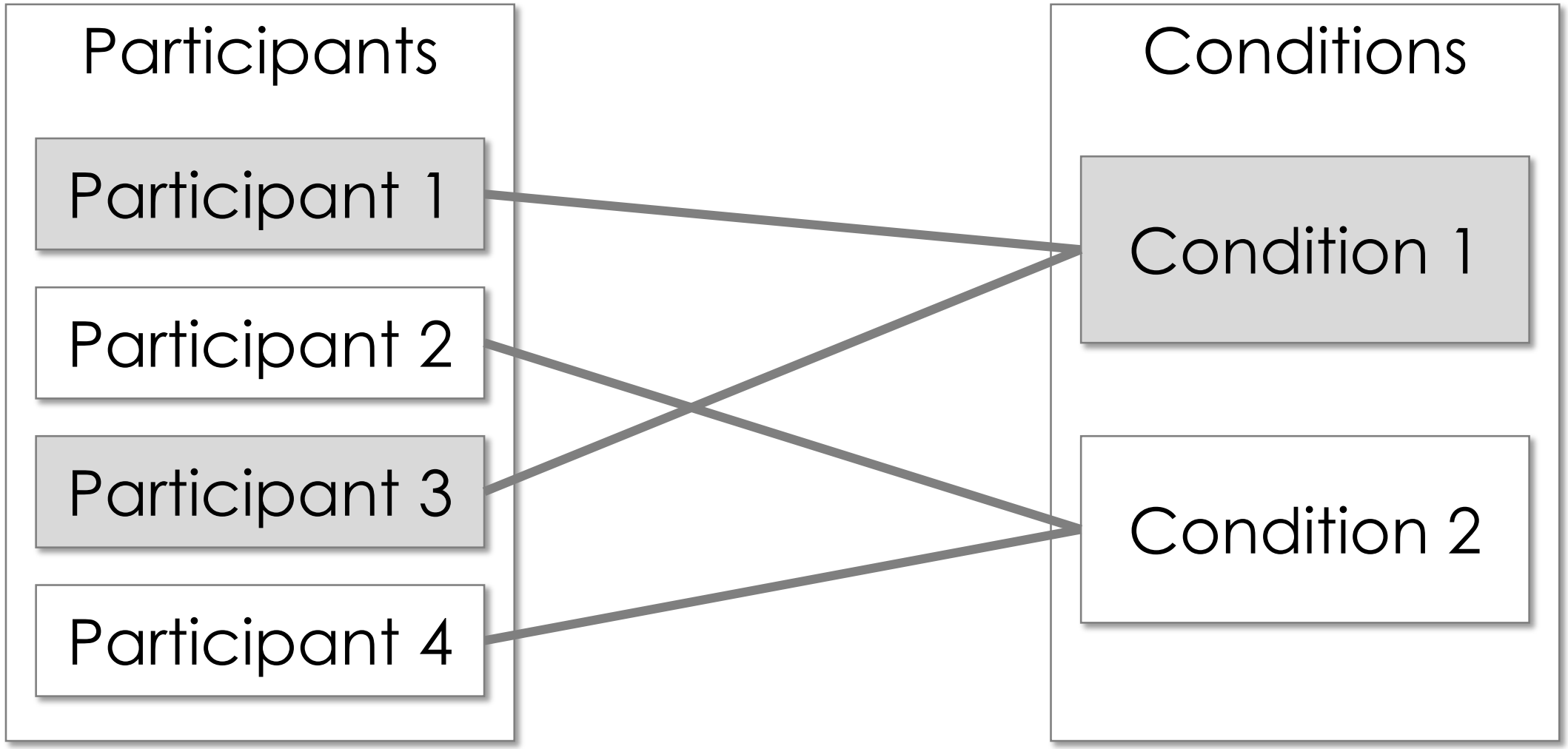


Within-participants design

Study 2 was a 2 (condition: e-mail vs. voice) x 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 SDs away from the mean on several dependent variables and were excluded from the analysis, yielding a final sample size of 29 dyads.

Kruger, J., Epley, N., Parker, J., and Ng, Z. (2005). Egocentrism over e-mail: Can we communicate as well as we think? Journal of Personality and Social Psychology, 89 (6): 925-936.

Between-participants design

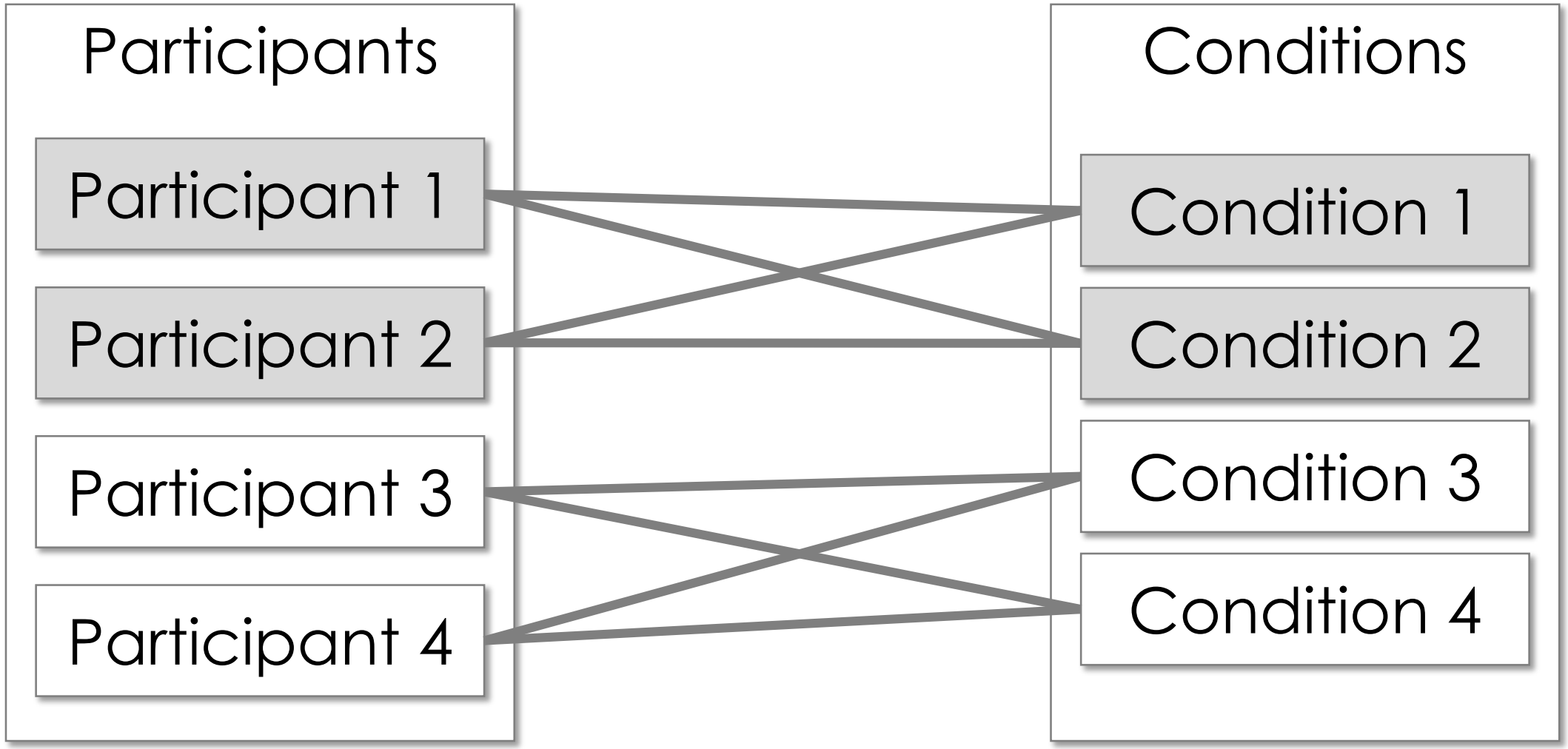


Between-participants design

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.

Hinds, P.J., Roberts, T.L., and Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. HumanComputer Interaction, 19, 151-181.

Mixed-model design



Mixed-model design

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) x 2 (order: Round 1 vs. Round 2) x 2 (acquaintanceship: stranger vs. friend) x 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.**

Kruger, J., Epley, N., Parker, J., and Ng, Z. (2005). Egocentrism over e-mail: Can we communicate as well as we think? Journal of Personality and Social Psychology, 89 (6): 925-936.

How to Choose Designs

Within-participants design – also called "repeated measures design"

- When *transfer effects** and *demand characteristics*† are not expected to be severe
  - * *Taking part in earlier trials changes performance in later trials due to learning, fatigue, etc.*
  - † *Participants trying to question the purpose of the experiment*
- When there are too many conditions that makes running the study infeasible (due to high number of participants required)
- When inter-participant variance is expected to be high
  - E.g., in performance-based measurements
  - Participants act as their own control

How to Choose Designs

Between-participants design

- When transfer/carryover effects and demand characteristics are a concern

- When the number of conditions is not too high and recruiting a high number of participants is feasible

- When inter-participant variance is not expected to be too high

How to Choose Designs

- Tradeoff between **power** and **bias**
  - Within-participants designs provide higher power
    - They control for the biggest source of variance in the data – variance across participants
      - Every subject serves as their own control
  - Within-participants designs suffer from sources of bias
    - Transfer or carryover effects – performance changes across trials due to learning, fatigue, etc.
    - Demand characteristics – participants start questioning the purpose of the experiment

How to Choose Designs

Within-participants design

Advantages

- Needs fewer participants
- Provides more statistical power

Disadvantages

- Might impose bias due to transfer effects, demand characteristics
- Difficult to administer – counterbalancing, etc.

How to Choose Designs

Between-participants design

Advantages

- Reduces bias by avoiding transfer effects and alleviating demand characteristics
- Easy to administer – you don't need to worry about counterbalancing, etc.

Disadvantages

- High variance in data due to inter-participant variability

How to Choose Designs

- Mixed-model design
  - Draws on the strengths of both designs in experiments with multiple factors (factorial designs)
  - Combine both designs when appropriate
  - Disadvantages
    - Difficult to administer, analyze, and interpret

Questions?

Step V:
Develop Experimental Task and Procedure

Experimental Task

A task context that represents real-world cognitive, social, or organizational situations

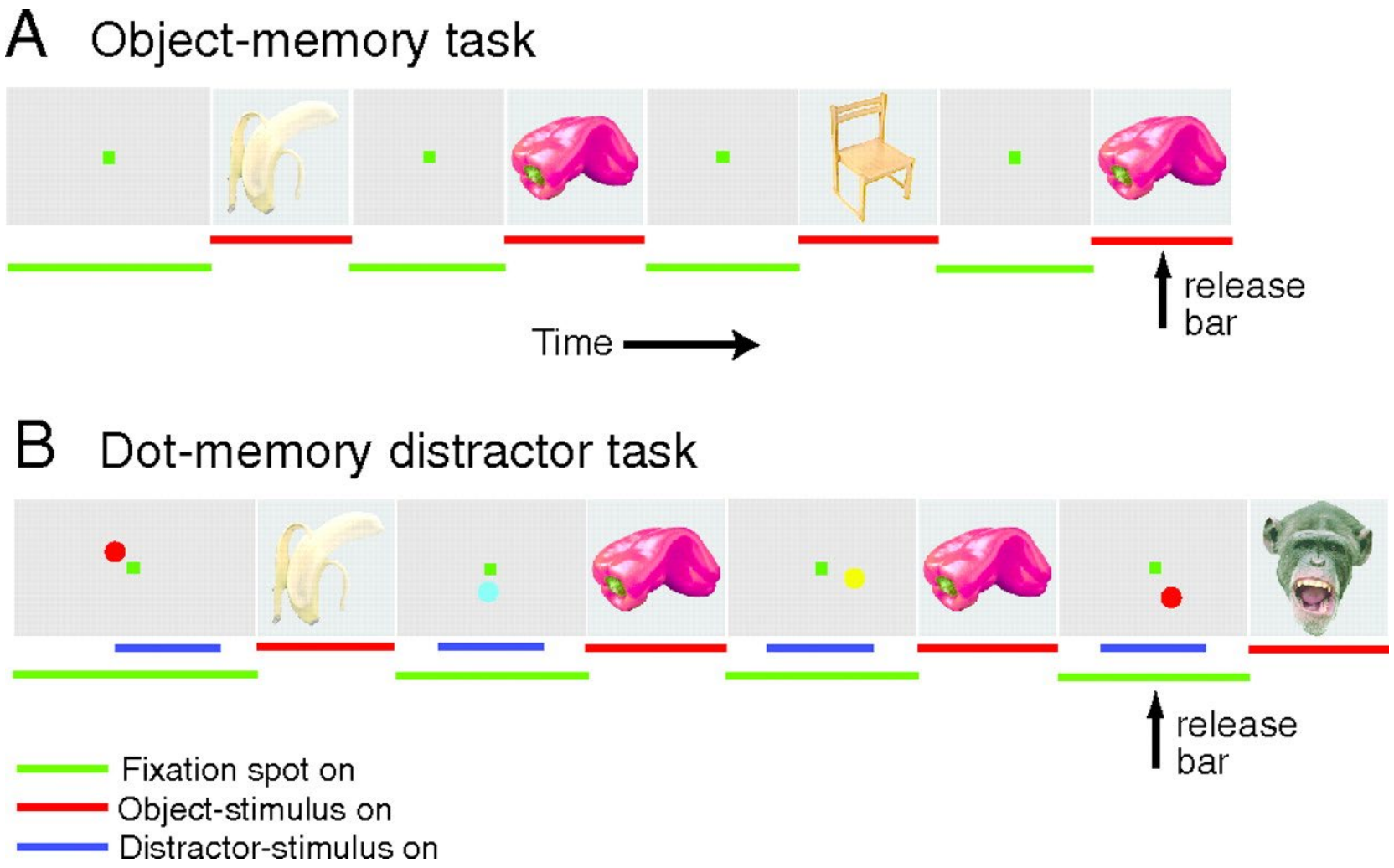
- Task should provide a reasonable context to test hypotheses
  - Relevant, reasonable, intuitive, easy-to-interpret (free from hard-to-control factors)
- You should be able to generalize the results from the lab task into the real world (ecological validity)
- The task should involve goals and measures of success
  - Should provide the appropriate motivational mechanisms for participants to perform the task as expected
- Can use tasks from prior experiments
  - Advantage: validated (?)
  - Disadvantage: what will you learn that is new?

Experimental Task

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.

Torrey, C., Powers, A., Marge, M., Fussell, S.R., and Kiesler, S. (2006). Effects of adaptive robot dialogue on information exchange and social relations. In Proceedings of the 1st ACM/IEEE International Conference on Human-robot interaction.

Distractor Tasks



Deceptive Tasks

2.3.4 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.

Bailenson, J.N., Blascovich, J., Beall, A.C., and Loomis, J.M. (2001). Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. Presence: Teleoperators & Virtual Environments, 10 (6), 583-598

Experimental Procedure

A detailed description of steps involved in the experiment

- Should be replicable
- Should provide enough detail about:
  - Task and instructions provided to the participant
  - Participants' involvement
  - Experimenters' actions
  - Equipment used
  - Points at which consent was obtained, measurements taken, and compensation provided

Experimental Procedure

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.

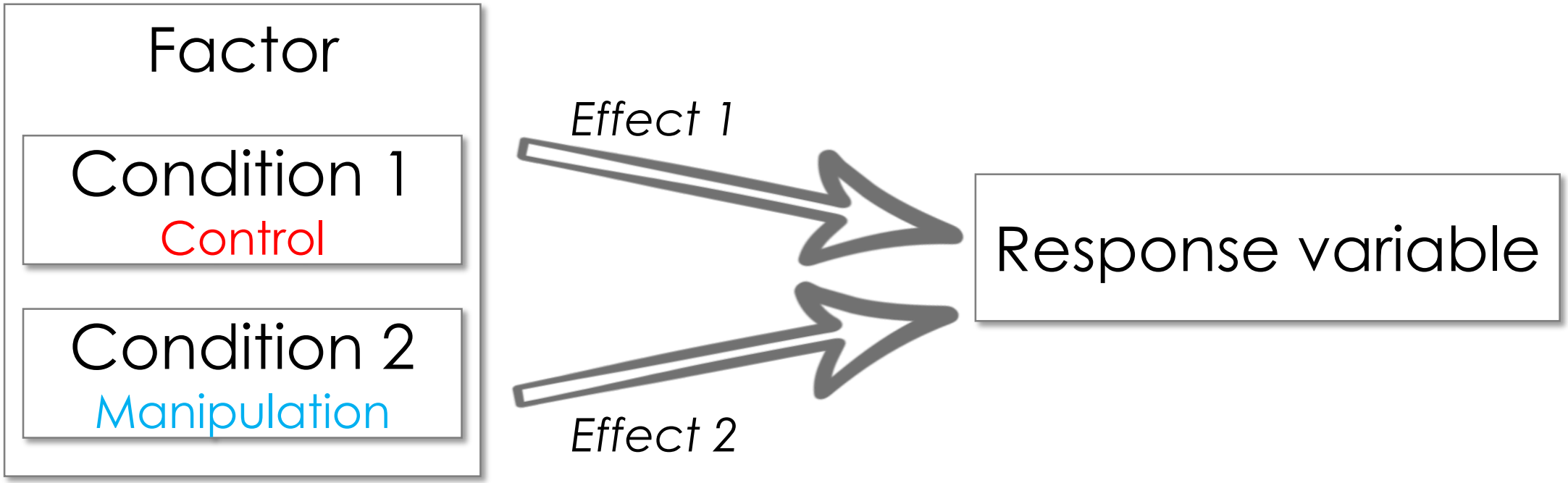
Torrey, C., Powers, A., Marge, M., Fussell, S.R., and Kiesler, S. (2006). Effects of adaptive robot dialogue on information exchange and social relations. In Proceedings of the 1st ACM/IEEE International Conference on Human-robot interaction

Questions?

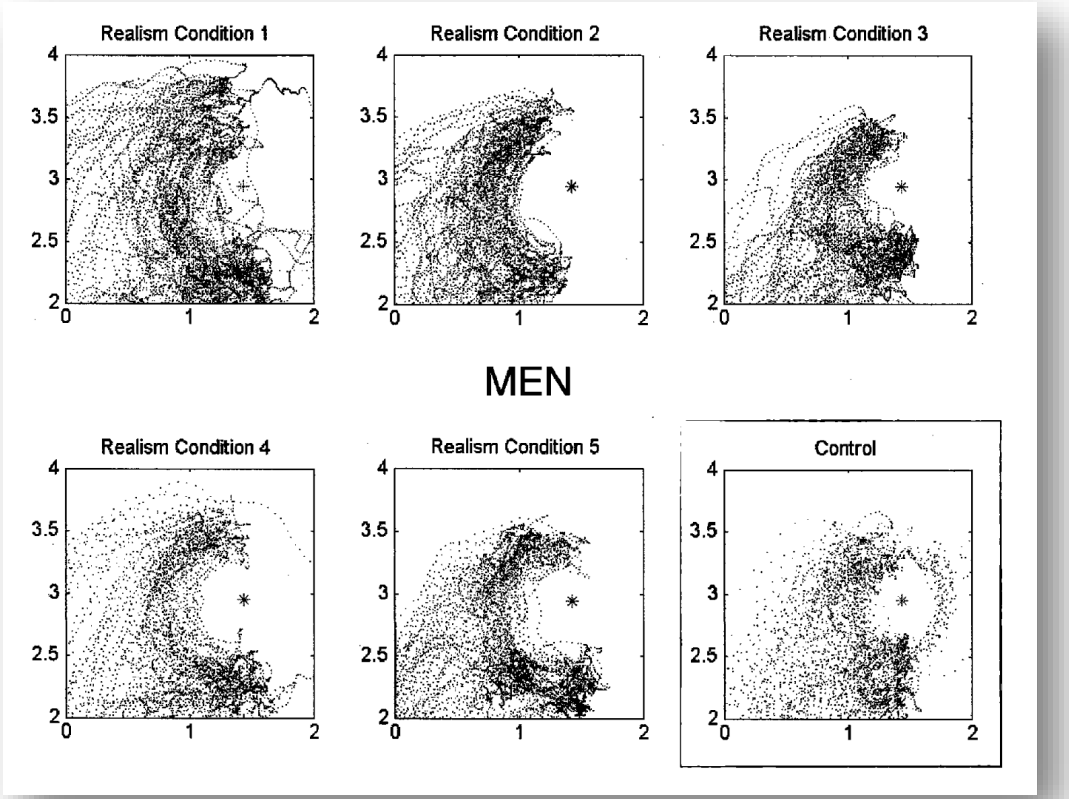
Step VI:
Determine Manipulations and Measurements

Manipulation

One level of the factor will be controlled, one will be manipulated



Control



Bailenson, J.N., Blascovich, J., Beall, A.C., and Loomis, J.M. (2001). Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. Presence: Teleoperators & Virtual Environments, 10 (6), 583-598

Measurements

Categorical, objective, subjective, behavioral

- Can be used as factors or response variables

More next class

Measurements

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).

Lee, K.M., Park, N., and Song, H. (2005). Can a robot be perceived as a developing creature? Effects of a Robot's Long-Term Cognitive Developments on Its Social Presence and People's Social Responses Toward It. Human Communication Research, 31 (4): 538-563

Questions?

Step VII:
Identify Participants

Participants

Has to be representative of the target population

Size has to provide statistical power

Balanced in measured factors preferred

Big source of bias – *needs to be carefully determined*

Participants

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.

Bailenson, J.N., Blascovich, J., Beall, A.C., and Loomis, J.M. (2001). Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. Presence: Teleoperators & Virtual Environments, 10 (6), 583-598.

Questions?

Assignment 2

Design an Experiment

Follow the steps to design an experiment and write up the "method" section of an experimental paper:

1. Formulate the research question – i.e., problem statement
2. Identify variables
3. Generate hypotheses
4. Determine experimental design
5. Develop experimental task and procedure
6. Determine manipulations and measurements
7. Identify participants

University of Colorado
Boulder

THANKS!

Professor **Dan Szafir**

Computer Science & ATLAS Institute
University of Colorado Boulder