MODULE 5: CONTROLLED EXPERIMENT DESIGN

Nicholas Ho, PhD

Institute of System Science, NUS

Pepper: A Human-Robot-Interaction Focused Robot

CITEC

Cognitive Interaction Technology
Cluster of Excellence
Bielefeld University

CITEC conducted research on human-robot interaction using Pepper (which is developed by SoftBank Robotics); to **study social interaction and how patterns of motion are learned**

- One 3-D and two HD cameras
- Two ultrasonic microphones and speakers
- Six laser sensors
- Four directional microphones
- Tablet computer as an input interface
- Three omni-directional wheels

Pepper: A Human-Robot-Interaction Focused Robot

Researchers at CITEC

- Transformed Pepper into a robot that is able to **reliably recognize its environment and attentively understand reactions from humans**
- Are especially **interested in the interface between human and robot**; include integrating AR system to allow people to view Pepper's status from an AR device (e.g. planning route, battery level)
- Taught Pepper to **throw a ball in a cup** and also **to be a museum guide that has to deal with customers' behavior**

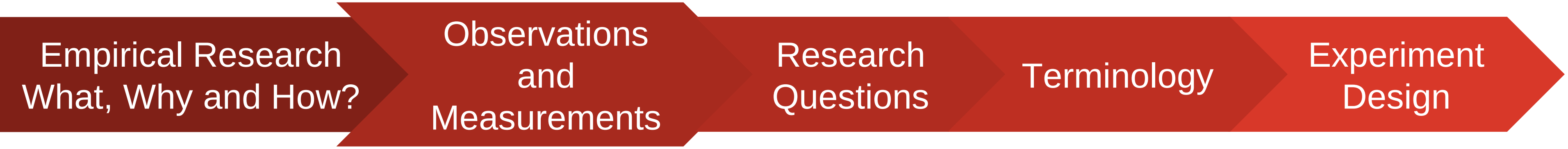
Pepper: A Human-Robot-Interaction Focused Robot

Universität Bielefeld
CITEC
With the help of us humans, Pepper learns complex movements.
Mit der Hilfe von uns Menschen lernt Pepper komplexe Bewegungen.

Source: https://www.youtube.com/watch?v=0cR26duOhDA

EMPIRICAL RESEARCH FOR HUMAN-ROBOT INTERACTION

Topics



Acknowledgement: these slides were adapted from Scott MacKenzie’s course in CHI and Shengdong Zhao’s workshop in NUS.

What is Empirical Research???

Empirical Research is ...

- Experimentation to discover and interpret **facts**, revise **theories** or **laws**
- Capable of being verified or disproved by observation or experiment

Why do Empirical Research???

We conduct empirical research to

- Answer (and raise!) questions about new or existing user interface designs or interaction techniques

- Find cause-and-effect relationships
- Transform baseless opinions into informed opinions supported by evidence
- Develop or test models that describe or predict behavior (of humans interacting with robots/computers)

How do we do Empirical Research???

We conduct empirical research through …

- a program of inquiry conforming to the scientific method

The scientific method involves …

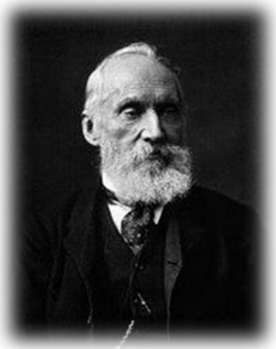
- The recognition and formulation of a problem
- The formulation and testing of hypotheses
- The collection of data through observation and experiment

Observe and Measure

Observations are gathered …

- Manually (human observers)
- Automatically (computers, software, cameras, sensors, etc)

A measurement is a recorded observation

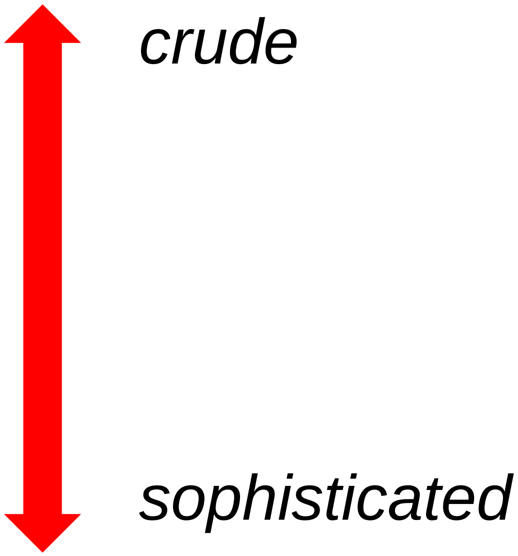


"When you cannot measure, your knowledge is of a meager and unsatisfactory kind."

--- William Thomson, 1st Baron Kelvin (1824 - 1907)

Scales of Measurement

- **Nominal**
- **Ordinal**
- **Interval**
- **Ratio**



crude

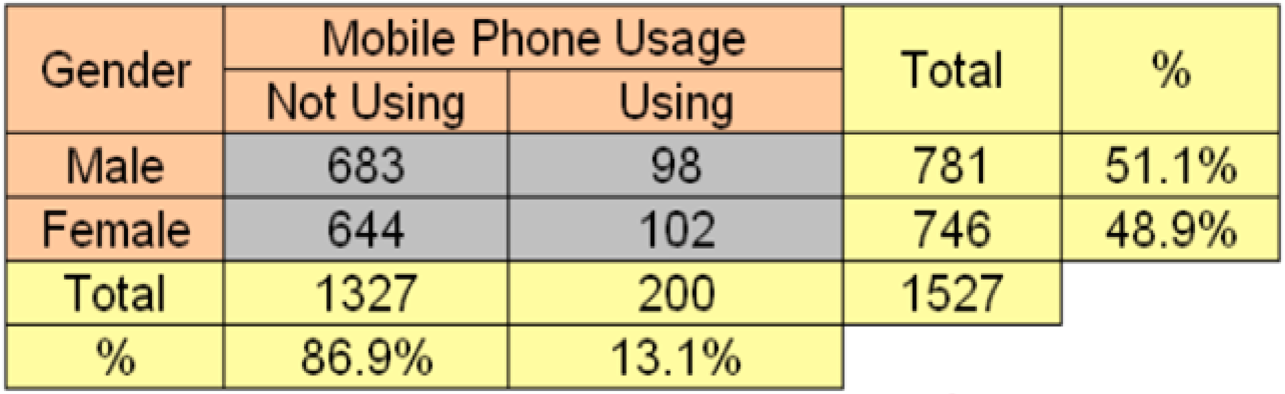
sophisticated

Nominal Data

- **Nominal data (a.k.a. categorical data) are arbitrary codes assigned to attributes**
  - M = male, F = female
  - 1 = mouse, 2 = touchpad, 3 = laser pen
- **Obviously, the statistical mean cannot be computed on nominal data**
- **Usually, it is the count that is important**
  - “Are females or males more likely to …”
  - “Do left or right handers have more difficulty with …”

Nominal Data Example

- **Observe students “on the move” on university campus**
- **Code and count students by …**
  - Gender (male, female)
  - Mobile phone usage (not using, using)



Gender	Mobile Phone Usage		Total	%
	Not Using	Using		
Male	683	98	781	51.1%
Female	644	102	746	48.9%
Total	1327	200	1527	
%	86.9%	13.1%		

Ordinal Data

- Ordinal data associate **order** or **rank** to an attribute
- The attribute is any characteristic or circumstance of interest
  - Users try 3 different GPS systems for a period of time
  - Then rank them: 1st, 2nd, 3rd choice
- More sophisticated than nominal data
  - Comparisons of “greater than” or “less than” possible

Ordinal Data Example

How many text messages do you send each day?

○ < 10

○ 10 - 50

○ 51 - 99

○ 100 - 200

○ > 200

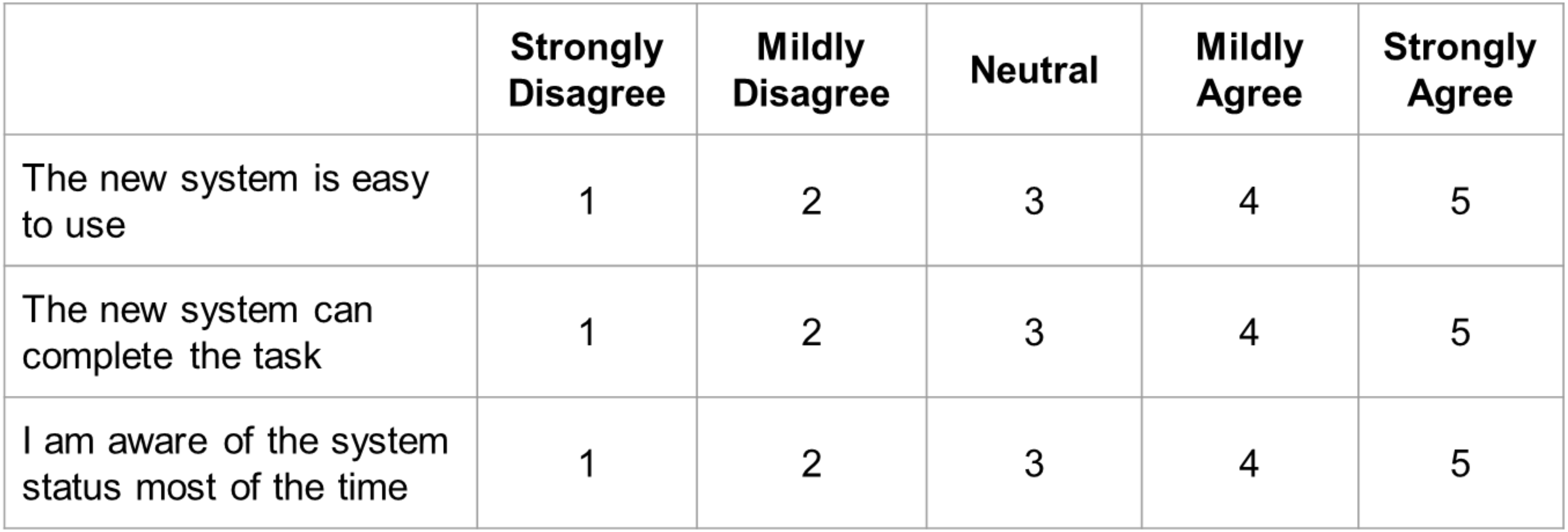
Interval Data

- **Equal distances between adjacent values**
- **But, no absolute zero**
- **Classic example: temperature (℉, ℃)**
- **Statistical mean possible**
  - The mean midday temperature during July
- **Ratio no possible**
  - Cannot say 10 ℃ is twice 5 ℃

Interval Data Example

- **Questionnaires often solicit a level of agreement to a statement**
- **Responses on a Likert scale**
- **Likert scale characteristics**
  - Statement soliciting level of agreement
  - Responses are symmetric about a neutral middle value
  - Gradations between responses are equal (more-or-less)
- **Assuming ”equal gradations”, the statistical mean is valid (and related statistical tests are possible)**

Interval Data Example (cont)

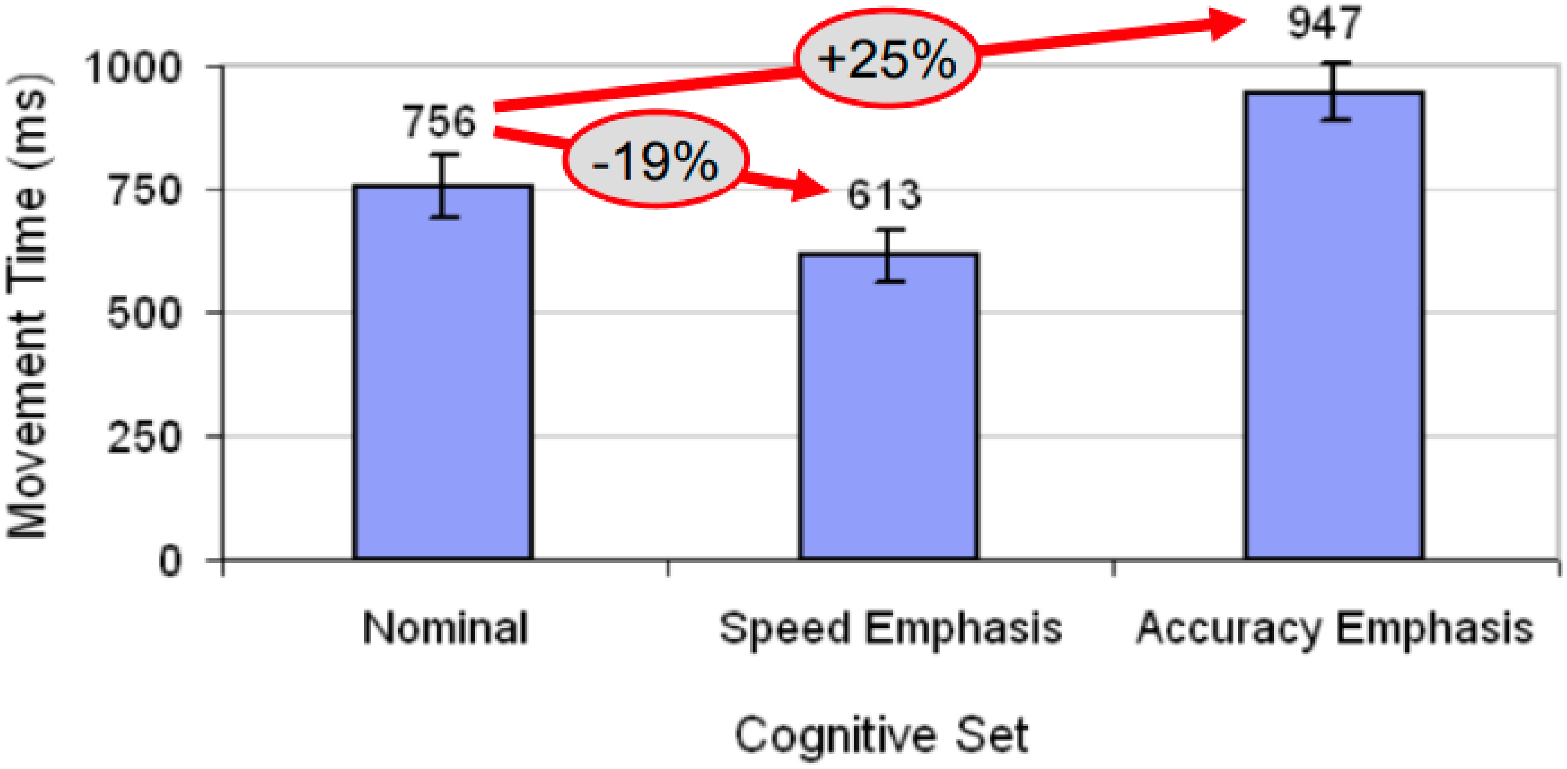


	Strongly Disagree	Mildly Disagree	Neutral	Mildly Agree	Strongly Agree
The new system is easy to use	1	2	3	4	5
The new system can complete the task	1	2	3	4	5
I am aware of the system status most of the time	1	2	3	4	5

Ratio Data

- Most sophisticated of the four scales of measurement
- Preferred scale of measurement
- Absolute zero, therefore many calculations possible
- Summaries and comparisons are strengthened
- A “count” is a ratio-scale measurement
  - Time (the number of seconds to complete a task)
- Enhance counts by adding further ratios where possible
  - Facilitates comparisons
  - E.g. A 10-word phrase was entered in 30 seconds
    - BAD: t = 0.5 minute
    - GOOD: Entry rate = 20 words-per-minute

Ratio Data Example



Research Questions

- **Consider the following questions**
  - Is it viable?
  - Is it better than current practice?
  - Which of the several design alternatives is the best?
  - What are the performance limits and capabilities?
  - What are the strengths and weaknesses?
  - Does it work well for novices, for experts?
  - How much practice is required to become proficient?
- **Are these good questions?**

Human-Robot Interaction
by The University of British Columbia

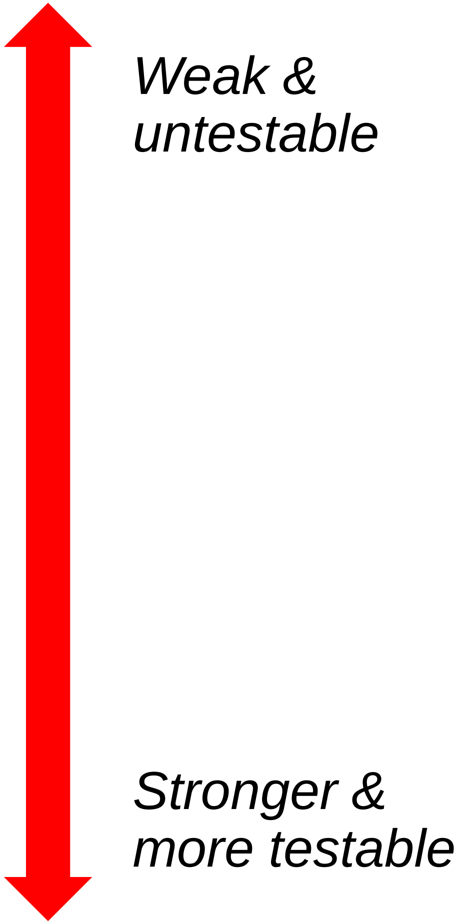
Source: https://www.youtube.com/watch?v=5AQ-E3njViw

Testable Research Questions

- Preceding questions, while unquestionably relevant, are **not testable**
- Try to re-cast as testable questions (even though the new question may appear less important)
- Scenario …
  - You have invented a new user interface for photo taking using flying cameras, and you think it is better than the existing joystick/joypad interface widely used today
  - You decide to undertake a program of empirical enquiry to evaluation your system
  - What are your research questions?

Testable Research Questions (cont'd)

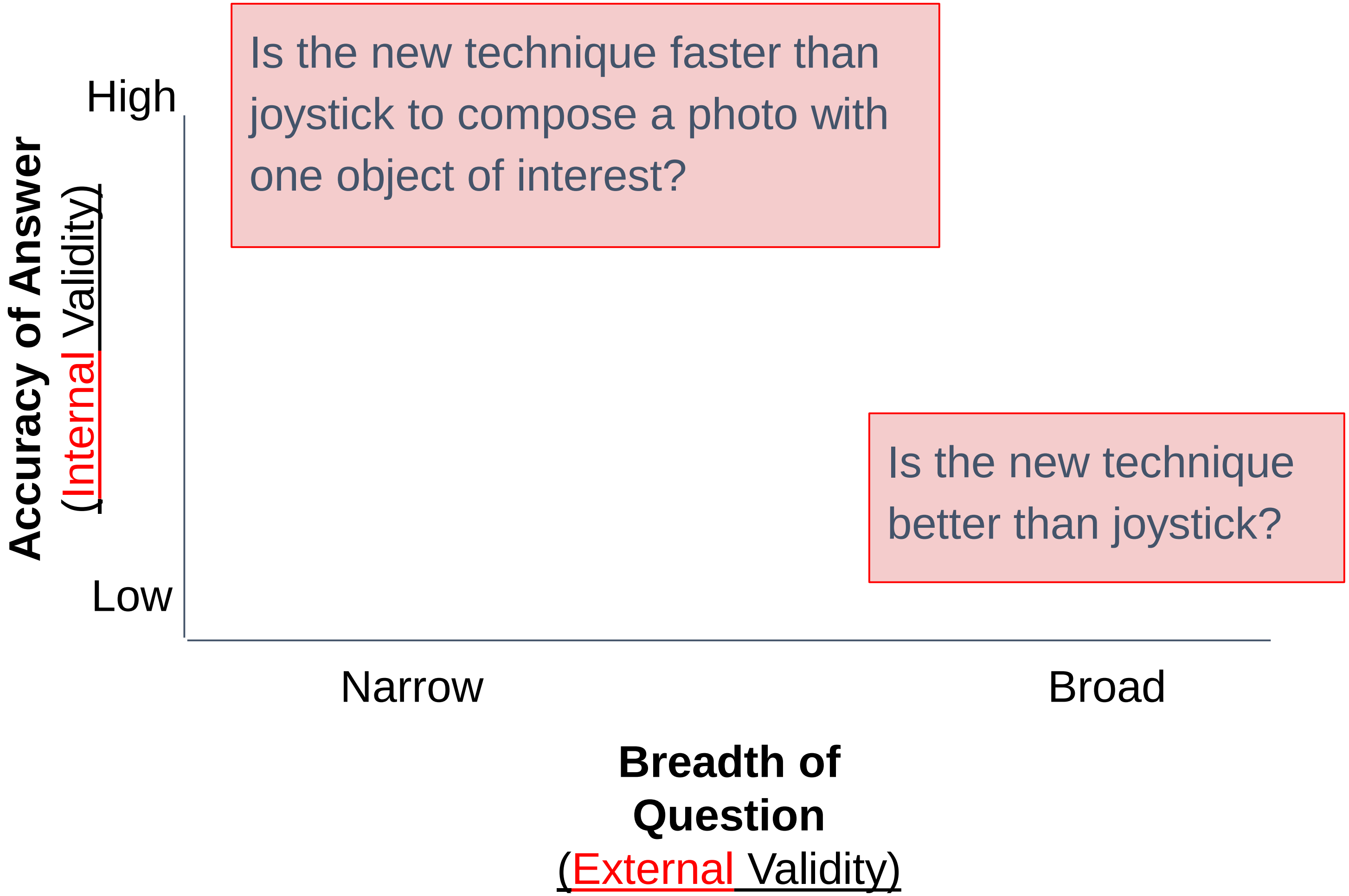
- Is the new technique any good?
- Is the new technique better than joystick?
- Is the new technique faster than joystick?
- Is the new technique faster than joystick to compose a photo with one object of interest?



Weak & untestable

Stronger & more testable

Tradeoff



Is the new technique faster than joystick to compose a photo with one object of interest?

Is the new technique better than joystick?

High

Low

Accuracy of Answer

(Internal Validity)

Narrow

Broad

Breadth of Question

(External Validity)

Internal Validity

- **Definition:**
  - The extent to which the effects observed are due to the test conditions
  - E.g. joystick vs new

- **Statistically …**
  - Differences (in the means) are due to inherent properties of the test conditions
  - Variances are due to participant differences
  - Other potential source of variance are controlled or exist equally and randomly across the test conditions

External Validity

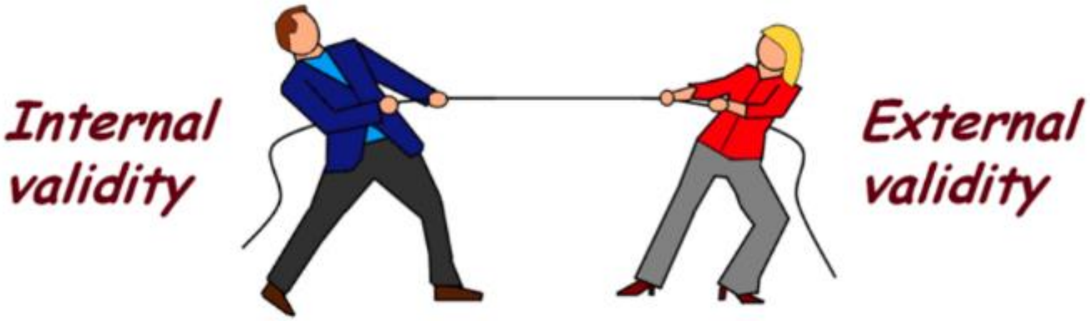
- **Definition:**
  - The extent to which the results are generalizable to other people and other situations

- **Statistically …**
  - The participants are representative of the broader intended population of users
  - The test environment and experiment tasks are representative of real world situations with the interface or technique will be used

Test Environment Example

- Scenario …
  - You wish to compare two interfaces for flying camera photography
- External validity is improved if the test environment mimics expected usage
- Test environment should probably involves ..
  - Taking selfie in a scenery place, e.g., the Marina Bay
  - Let participants use their own mobile devices
  - Let them take photos freely as they like
- But … is internal validity compromised?

Tradeoff



- There is tension between internal and external validity
- The more the test environment and experimental tasks are "relaxed" (to mimic real-world situations), the more the experiment is susceptible to **uncontrolled sources of variation**, such as environmental variations, distractions, or secondary tasks.
- How can we deal with the tradeoff??

Best of both worlds

- **Internal and external validity are increased by …**
  - Posing multiple narrow (testable) questions that cover the range of outcomes influencing the **broader** (untestable) questions
  - E.g., A technique that is faster, is more accurate, take fewer steps, is easier to learn, and is easier to remember, is generally **better**

- **The good news**
  - There is usually a positive correlation between the testable and untestable questions
    - Participants generally find a system **better** if it is faster, more accurate, takes fewer step, easier to learn and remember, etc

Common Terminology for Controlled Experiment Design

- **Participant**
- **Independent variable (a.k.a. factor)**
- **Test conditions (a.k.a. levels)**
- **Dependent variable**
- **Control variable**
- **Confounding variable**
- **Within subjects vs Between subjects**
- **Counterbalancing**
- **Latin square**

Participant

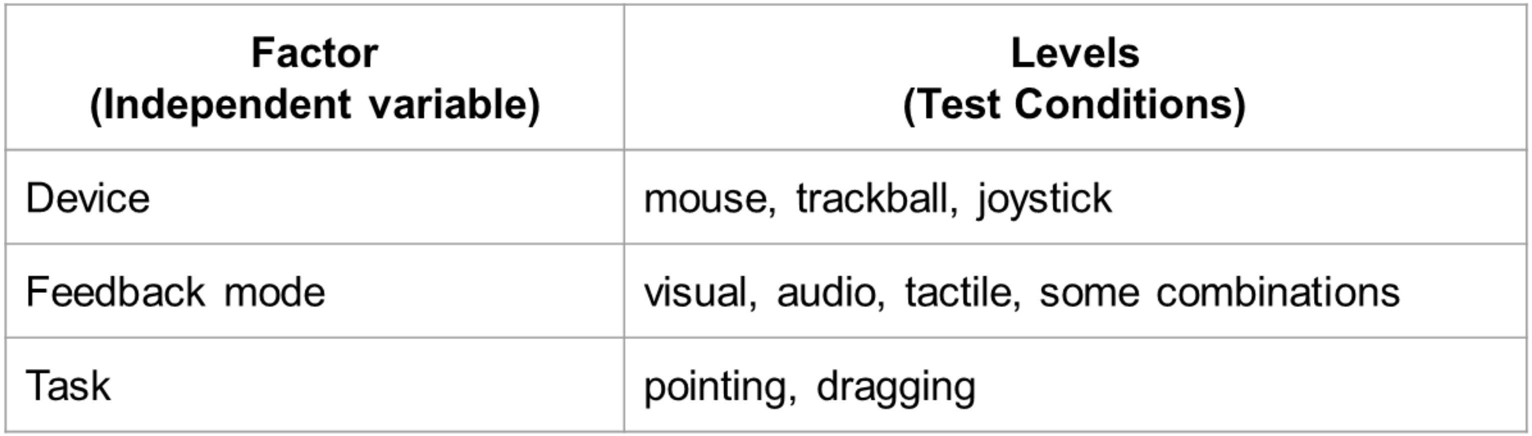
- The people participating in an experiment are referred to as participants
- Previously the term subjects was used, but it is no long in vogue
- When referring specifically to the experiment, use the term participants
  - E.g. “all participants exhibited a high error rate …”
- General comments on the problem or conclusion drawn may use other terms
  - E.g. “these results suggest that users are less likely to …”
- Report the selection criteria and give relevant demographic information or prior experience
  - E.g. “8 volunteers (2 female, 6 male, aged 23– 30) were recruited from the university community and the IT industry. All participants had prior experience taking photos, and 3 had experience flying drones.”

Independent variable (a.k.a. factor)

- An **independent variable** is a variable that is manipulated through the design of the experiment
- It is “independent” because it is independent of participant behaviour
  - there is nothing a participant can do to influence an independent variable
- E.g., interface, device, feedback mode, button layout, visual layout, gender, age, expertise, etc

Test conditions (a.k.a. levels)

- The level, values, or settings for an independent variable are the **test conditions**
- Provide a name for both the factor (independent variable) and its levels (test conditions)
- E.g.



Factor (Independent variable)	Levels (Test Conditions)
Device	mouse, trackball, joystick
Feedback mode	visual, audio, tactile, some combinations
Task	pointing, dragging

Dependent variable

- A **dependent variable** is a variable representing the measurements or observations on an independent variable
- E.g., task completion time, speed, accuracy, error rate, etc
- Give a name to the dependent variable, separate from its units
    - E.g. “Text entry speed” is a dependent variable with units “words per minute”

Control variable

- A **control variable** is a circumstance (not under investigation) that is kept constant to test the effect of an independent variable
- More control means the experiment is less generalizable, i.e. less applicable to other people and other situations
- E.g. room size, initial battery level, wind speed

Confounding variable

- A **confounding variable** is a circumstance that varies systematically with an independent variable
- It should be controlled or randomized to avoid misleading results
- E.g. 1, “Order”
  - All participants are tested on A, followed by B, followed by C
  - Performance might improve due to order (practice)
  - “Order” is a confounding variable
- E.g. 2, “Prior experience” (search engine interfaces)
  - All participants have prior experience with Google, but no experience with a new search engine
  - “Prior experience” is a confounding variable

Within Subjects, Between Subjects

- **Two ways to assign conditions to participants**
  - ○ Within-subjects: each participant is tested on each condition (a.k.a. repeated measures)
  - ○ Between-subjects: each participant is tested on one condition only

Participant	Test Condition		
1	A	B	C
2	A	B	C

Participant	Test Condition
1	A
2	A
3	B
4	B
5	C
6	C

Within Subjects, Between Subjects (cont)

	Within Subjects	Between Subjects
# participants	Fewer, easier to recruit, schedule, etc	More, harder to recruit, schedule, etc
Variation due to participants	Less	More
Balance groups	No need	Need to ensure the groups are more or less the same
Order effects	Interference between test conditions	No interference between test conditions

Counterbalancing

- For within-subjects designs, participants may benefit from the first condition and consequently perform better on the second condition - **we don’t want this!**
- To compensate, the order of presenting conditions is **counterbalanced**
- Participants are divided into ***groups***, and a different order of administration is used for each group
- The order is best governed by a **Latin Square** (next slide)

Latin Square

- The defining characteristic of a **Latin Square** is that each condition occurs only once in each row and column
- E.g.

A	B	C
B	C	A
C	A	B

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

A	B	C	D
B	D	A	C
D	C	B	A
C	A	D	B

Note: In a **Balanced Latin Square** each condition both precedes and follows each other condition an equal number of times

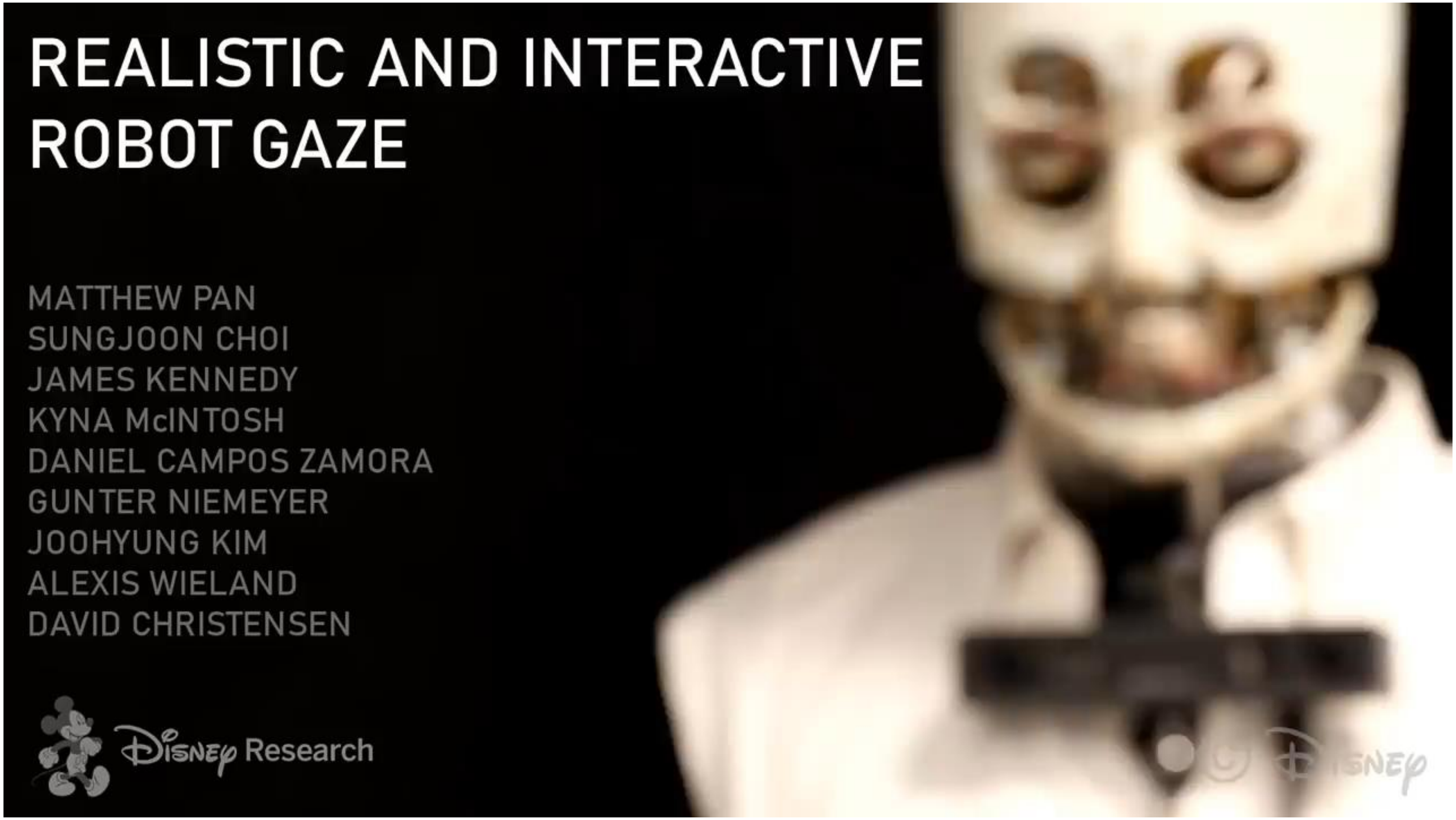
The Future of Human Robot Interactions by Accenture Technology

Source: https://www.youtube.com/watch?v=8CfRLTk8wpw

The Future of Human Robot Interactions by MIT CSAIL (Computer Science & Artificial Intelligence Laboratory)

Source: https://www.youtube.com/watch?v=Zd9WhJPa2Ok

The Future of Human Robot Interactions by Disney Research



REALISTIC AND INTERACTIVE ROBOT GAZE

MATTHEW PAN
SUNGJOON CHOI
JAMES KENNEDY
KYNA McINTOSH
DANIEL CAMPOS ZAMORA
GUNTER NIEMEYER
JOOHYUNG KIM
ALEXIS WIELAND
DAVID CHRISTENSEN

Source: https://www.youtube.com/watch?v=D8_VmWWRJgE&feature=youtu.be

THANK YOU

Email: nicholas.ho@nus.edu.sg