CMPT 363: User Interface Design
Summer 2021

Week 12: Evaluating Interfaces with Users: Experiments

Instructor: Victor Cheung, PhD

School of Computing Science, Simon Fraser University

Recap from Last Lecture

- Inclusive Design
  - Does not always mean disability, can simply be different
  - Diversity not "one-size-fits-all/most"
- Accessibility
  - Disabilities – different kinds lead to different barriers & needs
  - Assistive technologies – different levels of tech, found in modern operating systems

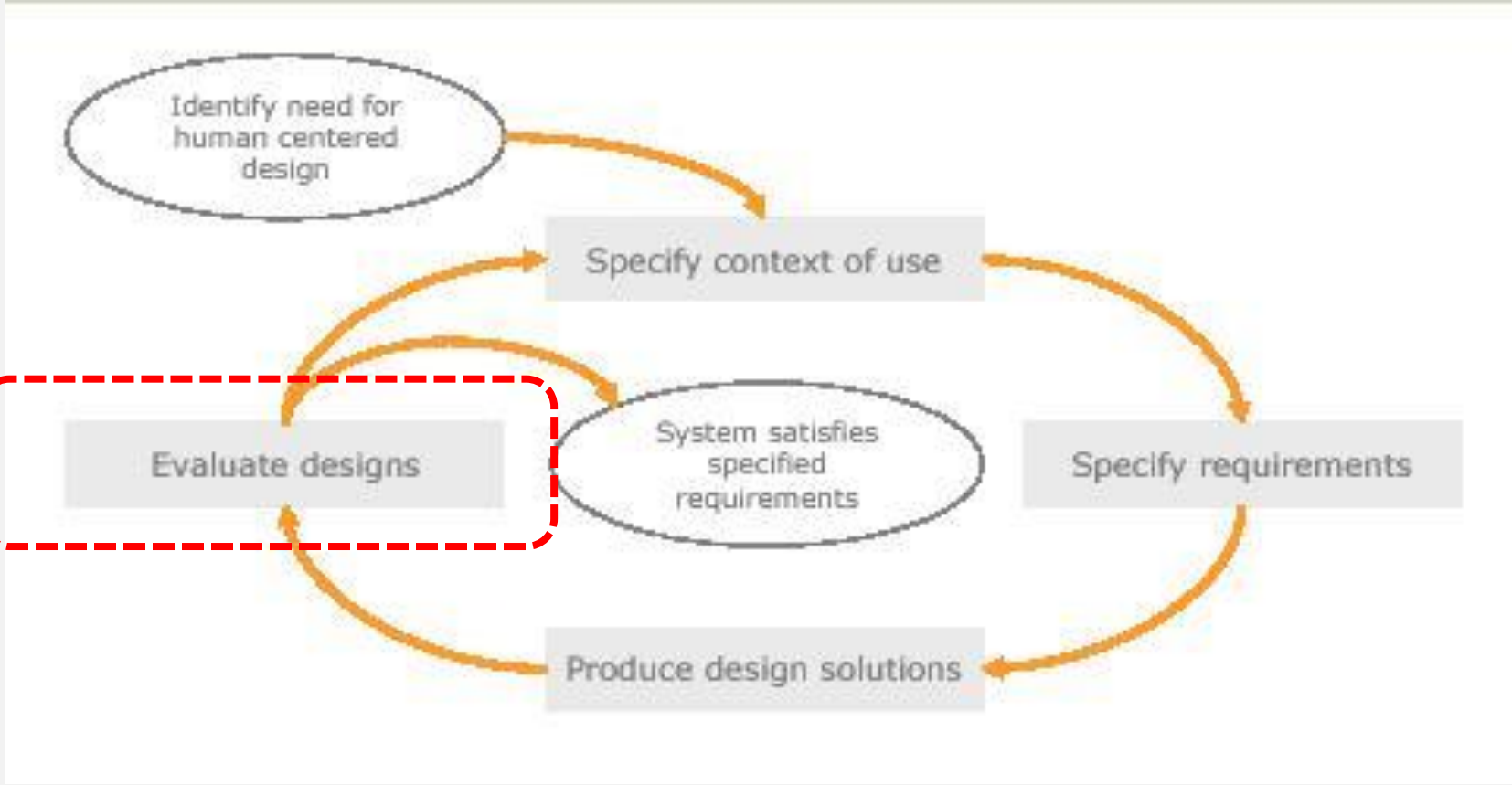
Group Project Part 3

- Overview
  - To design the interface for an online calendar that facilitates different kinds of activities for university students
- Part 3 (due on Aug 6) (https://canvas.sfu.ca/courses/63144/assignments/653608)
  - Continue with your MFPs
    - Cognitive Walkthrough
    - Reflection
    - Video demo (upload to SFU Vault by Aug 1)
- Group Project Contribution Form (individual) (due on Aug 9)

Today

- Evaluating Interfaces with Users: Experiments
  - Lab experiments
    - Why, what, where, who, how
    - Terminologies

The Human-Centered Design Process



ISO 13407:1999

Involving Users in Every Step

- Understand/specify context of use
  - Interview users & examine tasks
- Specify requirements
  - Verify & prioritize with users

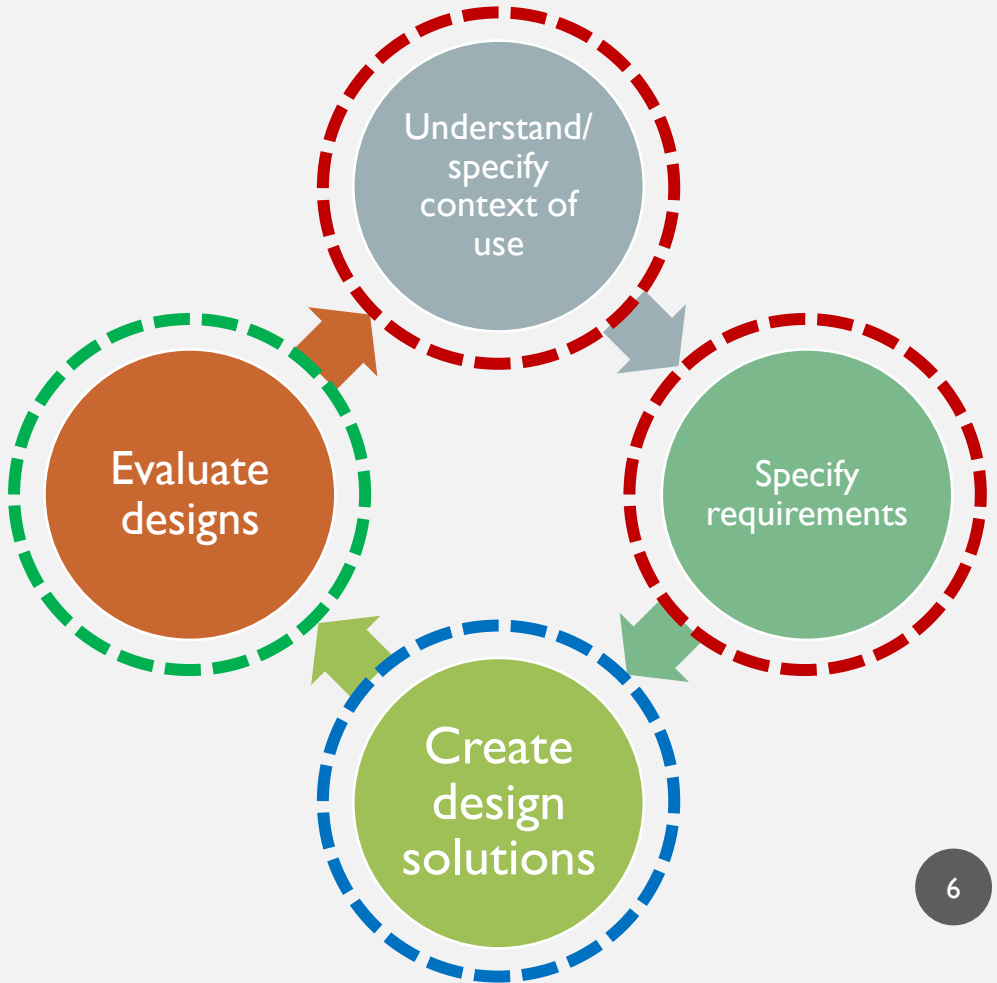
Week 4

- Create design solutions
  - Design with users (co-design)

Week 5

- Evaluate designs
  - Invite users to assess

Week 2 & 10

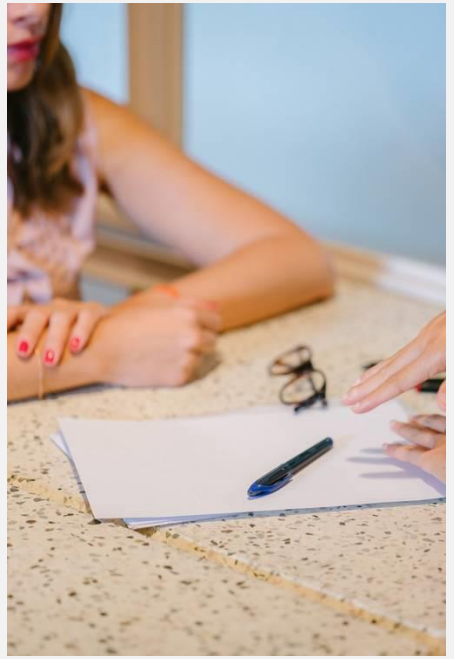


(Week 2 & 10) Types of Evaluation (ID-Book Ch 14.3)

- Controlled settings directly involving users (Week 2 & this lecture)
  - Usually done in labs to provide the most control (mostly called usability testing/studies/experiments)
- Natural settings involving users (Not covered)
  - Usually done outside labs where the interface is designed to be used at (mostly called in-the-wild studies)
- Any settings not directly involving users (Week 2 & 10)
  - Consultants/field experts instead of users (mostly called analytical evaluation)
    - Heuristic Evaluation
    - Cognitive Walkthrough
    - Fitts' Law, GOMS, KLM, …etc.

Controlled Settings Directly Involving Users

- Different names to this evaluation activity
  - Usability testing – because you are testing the usability attributes of a user interface
  - Usability studies – because you are studying how people use a user interface & evaluate usability
  - Usability experiments – because you are experimenting instances of a user interface
- HCI research often adopts protocols from psychology and medicine to evaluate
  - Psychology studies cause-effect relationships between stimulus and human responses
  - Medicine studies cause-effect relationships between administering of medication



Laboratory (Lab) Experiments

- Deductive reasoning
  1. Begin with a hypothesis (e.g., based on some theory, have something to compare)
  2. Try out possible variations of interest & make observations
  3. Use findings to confirm or reject hypothesis
- It is "controlled" because you want to isolate the cause and be sure that any effects are indeed due to that cause
  - Example: someone changes their diet & sleeping hours performed better in exam, but we cannot tell if it is the diet or the sleeping hours (can be just one, or both) → need a way to isolate diet & sleeping hours

Lab Experiments: Why

- The goal is to determine cause-effect relationships (like psychology & medicine)
  - "How does a change [in the interface] impact the participants' behaviour/performance?"
- An experimental process (protocol) provides a framework for establishing these cause-effect relationships

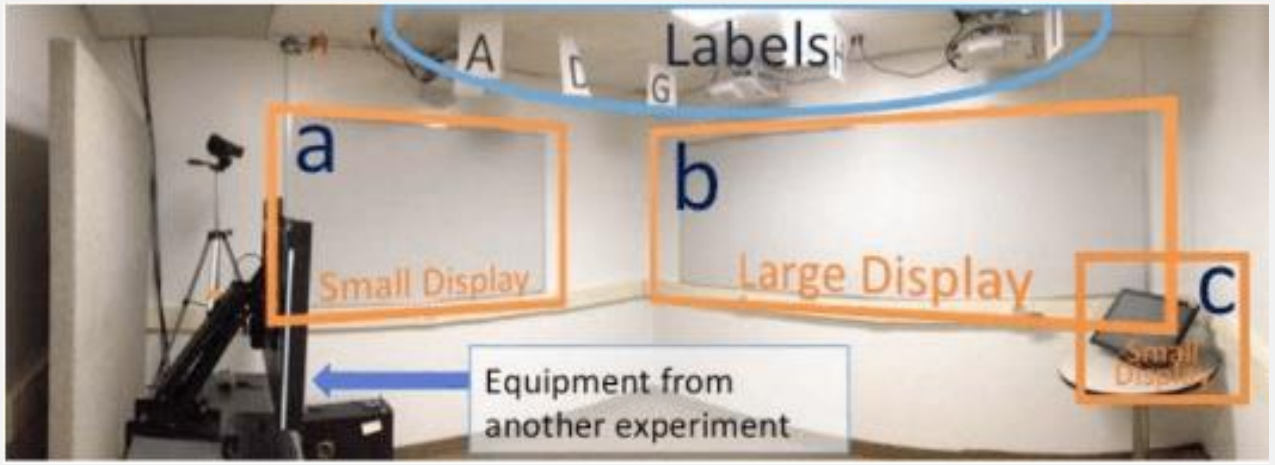
Lab Experiment: What

- Carefully measure performance on prepared tasks typical for interface's intended users
  - Tasks for each participants should be the same
- Make tightly controlled comparisons to isolate factors that may affect users
  - Remove as many external factors as possible (e.g., time of day/week, distractions)
- Focus on collecting quantitative data
  - Collected using a variety of tools: video/audio, screen-captures, keystrokes, notes
  - Typically measured as time of completion, error rate, accuracy, …etc., anything that can be collected objectively
  - Can also be subjective measurements such as confidence in using, preference level, …etc., anything that can be assigned to a numerical value (e.g., Likert Scale)
    - In HCI, qualitative data are also often collect in the form of observation and participant feedback/comments

Lab Experiments: Where

- Performed in laboratory-like conditions
  - To remove as many distractions as possible
  - To allow for equipment in obtaining quantitative measurements as precisely as possible

http://iat.ubalt.edu/usability_lab/



Cheung V. & Scott S.D. 2015. A laboratory-based study methodology to investigate attraction power of large public interactive displays. In Proceedings of UbiComp '15. ACM, New York, NY, USA, 1239–1250. DOI:https://doi.org/10.1145/2750858.2805842

Lab Experiment: Who

Gives instructions, answers participant's questions, observe and ask follow up questions.

Facilitator

Guides the participant through the test process

Tasks

Realistic activities that the participant might actually perform in real life

Participant

Realistic user of the product or service being studied

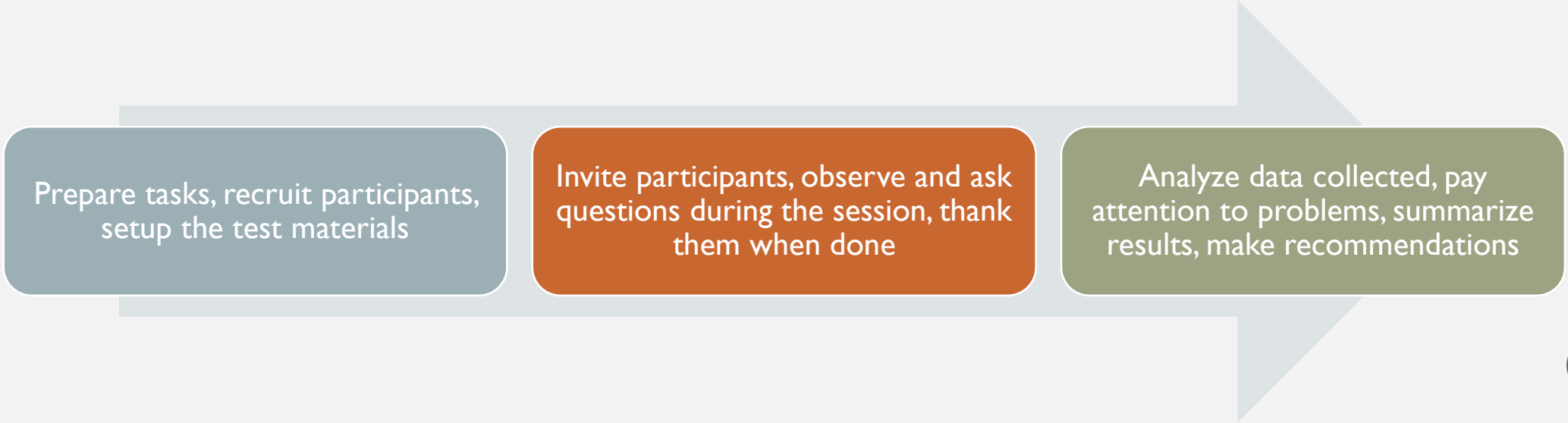
Someone who is already a user, or has a similar background and needs as a target user.

NNGROUP.COM NN/g

Source: https://www.nngroup.com/articles/usability-testing-101/

Lab Experiments: How

- Very formal process
  - Designed & piloted (to ensure all bases are covered, no ambiguity, and all parts are working)
  - Scripts ready for the facilitator to ensure consistency across all sessions
  - Obtain ethics approval (a required step in any academic settings and academic publication venues)



Pros & Cons of Lab Experiments

- Pros
  - Quantitative data lend themselves to statistical analysis (a powerful tool to reveal effects)
  - Allows focus on very specific questions
  - High on precision of measurements
- Cons
  - Significant setup costs (planning, pilot testing, participant recruitment…etc.)
  - Controlled environment strips out "reality" – lose the influence of peripheral, real-world activities, which might be important
  - Often needs a functional system for participants to use (even Wizard of Oz has limits)
  - Care must be taken to develop authentic & representative tasks, which might take too long to complete

Lab Experiments Terminology

- Overview
  - Hypothesis – a suggested explanation of a phenomenon based on observation or existing theories
  - Independent/dependent variables – values or settings that change in the experiment
  - Relationships – impact of independent variables on dependent variables
  - Conditions – situations created by changing independent variables
  - Between-/within-subjects – ways of participants exposed to the experiment conditions
  - Confounds, validity, reliability – things to check when planning, analyzing, & reporting results

Hypothesis

- A suggested explanation of a phenomenon
  - Example: "*If I change A, then B will change as the result*"
- Be as specific as possible so it is testable (and repeatable), i.e., can we test it to figure if it is true or false?
  - Non-testable example 1: Conscious is a result of illusion (we cannot measure consciousness)
  - Non-testable example 2: touch-screens are better than mouse (too broad to test, and they both excel in different aspects)
- A related term to "testable" is "fasifiable", i.e., can we disprove the hypothesis
  - An important attribute in science (if you come up with a theory, others can prove that it is wrong)
  - Example: "all swans are white" (you can disprove it by finding a black swan) vs "some pigs can fly" (you cannot disprove it because you cannot find all pigs that have or will have existed and show they cannot fly)

Exercise

- Poll – Which of the following statement is not testable/falsifiable?
  - A – There are no black swans
  - B – Aliens do not exist
  - C – There were more sunny days than cloudy days last year
  - D – The sun rises everyday

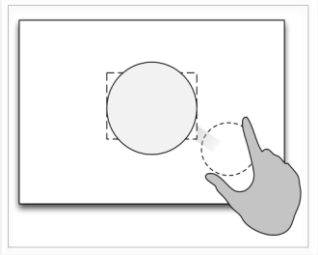
5min + 5min Break

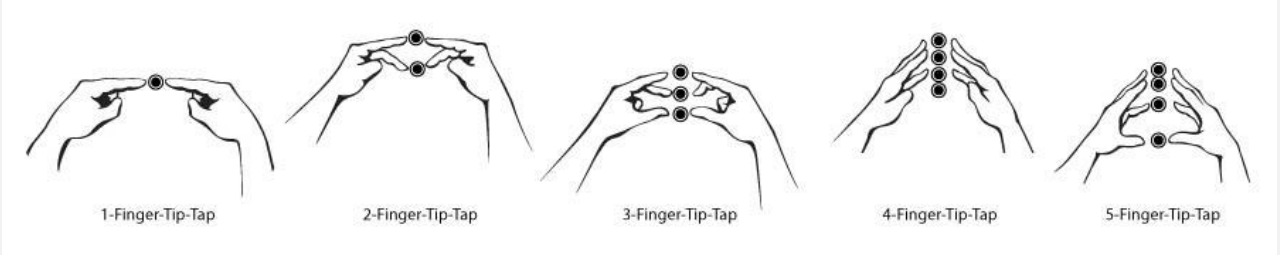
- Suppose there is a new interface for sending emails. Come up with:
  - 1 testable statement that can be stated as a hypothesis
  - 1 untestable statement

- Think about this: can we really test the testable statement? How sure can we be?

Typical Example of Hypothesis in HCI

- Hypothesis:
  "Using bimanual gestures, users will be able to more quickly and precisely zoom in and out on a tablet of a given size than when using one hand (unimanual)"
  - Makes the experimental variables clear
    - Experimenting bimanual gestures vs unimanual gestures
  - Makes the expected outcome clear
    - Ability to zoom in and out quickly and precisely on a tablet of a given size
- It is falsifiable because you can run an experiment to show it actually takes more time and causes more mistakes



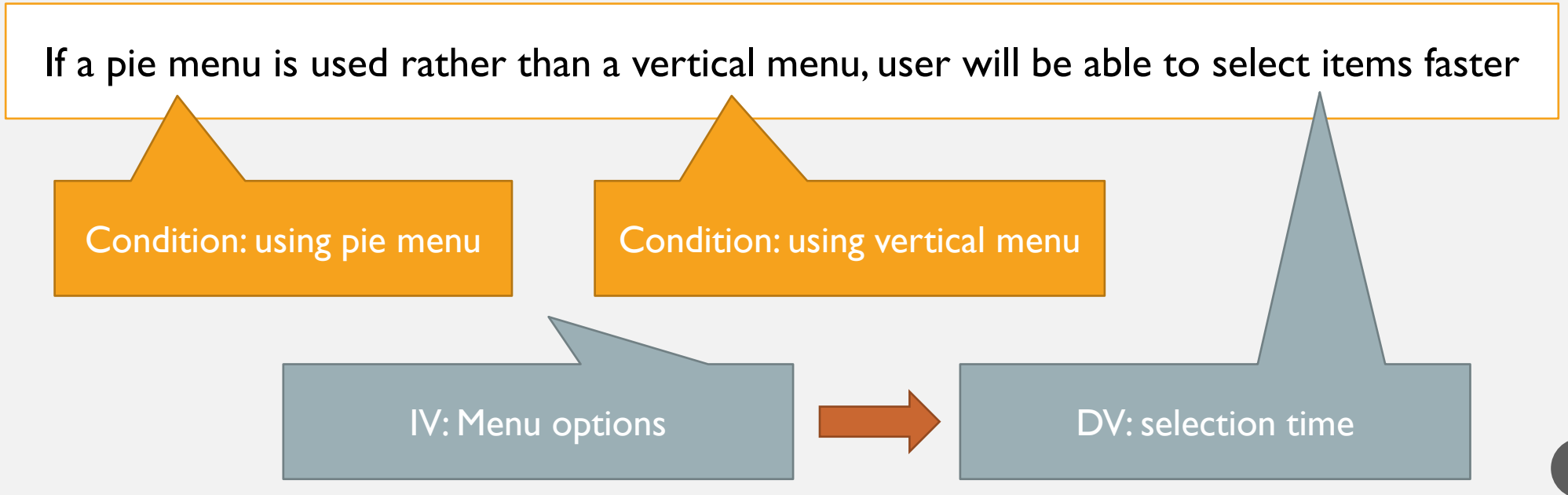


Variables

- Values or settings that change in the experiment
  - Independent – what you manipulate in the experiment (while holding everything else constant)
    - Examples: bimanual gestures vs unimanual gestures, standing vs sitting, big buttons vs small buttons
  - Dependent – what you expect to change
    - Examples: speed & precision, level of fatigue
    - Other examples: cognitive load, learning time, satisfaction & preference, talking time

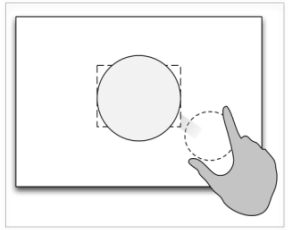
Relationships

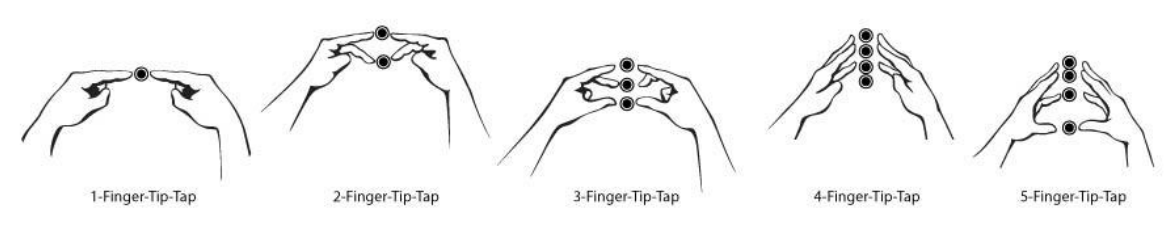
- Impact of independent variables on dependent variables
  - When being manipulated, independent variables are assumed to produce an effect on dependent variables' values

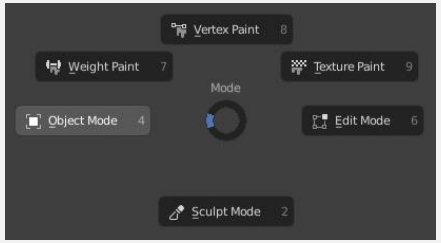


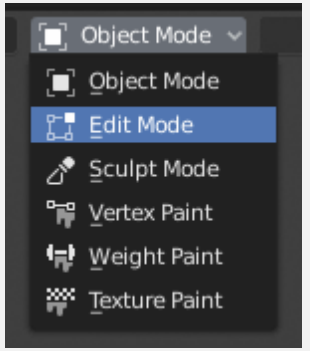
Conditions

- Situations created by changing independent variables
  - Most of the time independent variables are discrete (e.g., bimanual vs unimanual, shortcuts vs no shortcuts)
  - Each value that an independent variable takes is a condition (e.g., pie menu vs vertical menu = 2 conditions)
- Also referred to "levels" (not necessarily have the meaning of high/low)









Source: https://docs.blender.org/manual/en/latest/interface/controls/buttons/menus.html

Between-/Within-Subjects

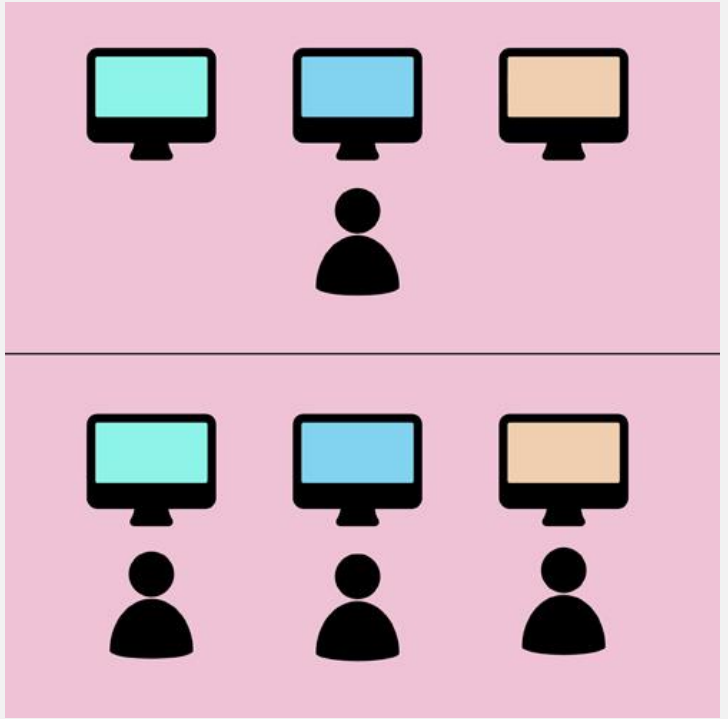
- Ways of participants exposed to the experiment conditions (also good to replace "subjects" with "participants")
  - Between-subjects – each participant only sees one condition
    - Pros: no learning effect (response won't be affected by previous condition), takes less time (only one condition each)
    - Cons: variations in participants (one person might just be better/worse in a condition), need more people to be valid
  - Within-subjects – every participant sees every condition
    - Pros: needs fewer participants, no trait variations, can also get insights on how they compare conditions
    - Cons: must be conducted carefully to eliminate learning effect (counter-balancing), takes longer

Covers all possible orders of conditions. E.g., with conditions A&B, half of the participants see A then B, the other half see B then A

Notes on Between-/Within-Subjects Experiments

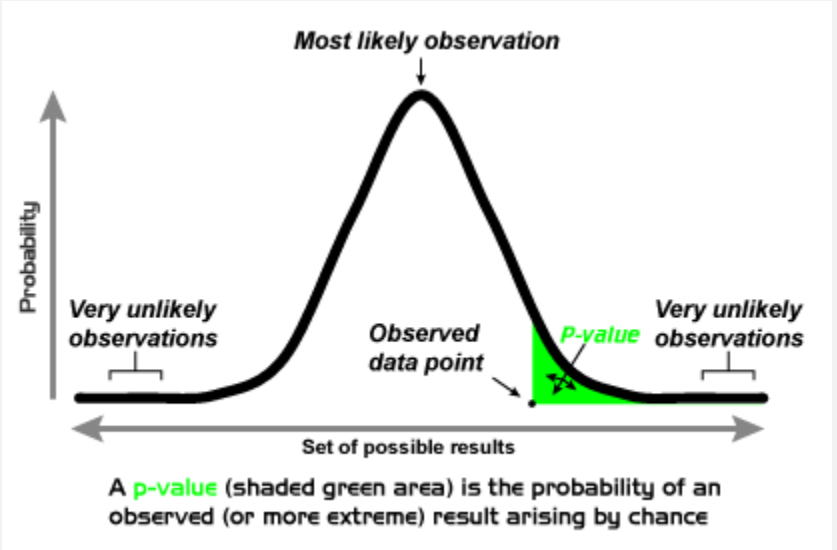
- Within-subjects experiments are generally preferred but it doesn't always work/viable:
  - Each condition takes a long time
  - Experiencing a condition will interfere the other significantly
    - Testing learnability
    - Strong learning effect

- Read more on https://measuringu.com/between-within/



Statistics Tools

- Descriptive statistics (measures of central tendency, variability, & trend)
  - Mean, median, mode
  - Standard deviation, variance, min/max
- Analytical statistics
  - Hypothesis testing
    - t-tests, Analysis of Variance, …etc.
  - Correlation, interaction



A one-tailed t-test for statistical significance

Hypothesis Testing

- In testing a hypothesis, we are seeking to disprove the null hypothesis
  - Null hypothesis: there exists no relationship between manipulating the independent variables and the resultant changes in the dependent variables (i.e., no affect, no relationship, between independent and dependent variables)

- For example, for the "bimanual vs unimanual gestures" hypothesis, we are trying to disprove the following:
  - There is no difference between bimanual & unimanual gestures in terms of users' speed and precision when users zoom in & out on a tablet of a given size
  - Using statics tools like t-tests, you are disproving a null hypothesis with a level of confidence (e.g., less than 5% of chance the difference happens by chance). This is because you are only running your experiment with a sample population

How Sure Are We Actually?

- We typically cannot have all possible users to participate a lab experiment (limited by budget, time, …etc.)
  - Say we find 30 participants and detected some impact of the IV on the DV, what about the 31st that we haven't invited?
- The participants we invite is a sample of the population we want to design for
  - So while we can do our best to find a representative group, we can't guarantee results from them are 100% absolute

- All statistic tools only give us an "estimate" or a "most likely to be true" result
  - descriptive statistics tell us an estimate
  - statistical analyses tell us a case where it is likely to be true (a small % of chance that it is actually not)

Confounds, Validity, Reliability

- Things to check when planning, analyzing, & reporting results
  - Confounds – Existence of variables/factors that wasn't controlled but might influence the results
    - E.g., someone who hasn't eaten yet vs someone who just had a big meal
  - Validity – Degree of accuracy to which the experiment measures what it is supposed to measure
    - E.g., measurement of precision is correct and is indeed caused by different conditions, result is also true for similar people
  - Reliability – Degree of consistency of a measurement
    - E.g., conducting the same experiment again will produce similar results

Going beyond Lab Experiments

- 8 main "research strategies"
  - Field experiments
  - Field studies
  - Experimental simulations
  - Laboratory experiments
  - Judgement tasks
  - Sample surveys
  - Formal theory
  - Computer simulations
- Concerned with Generality/Precision/Realism, but can't have all

Precision: measurement of behaviours

Realism: close to actual situation

Generality: applicable to the population

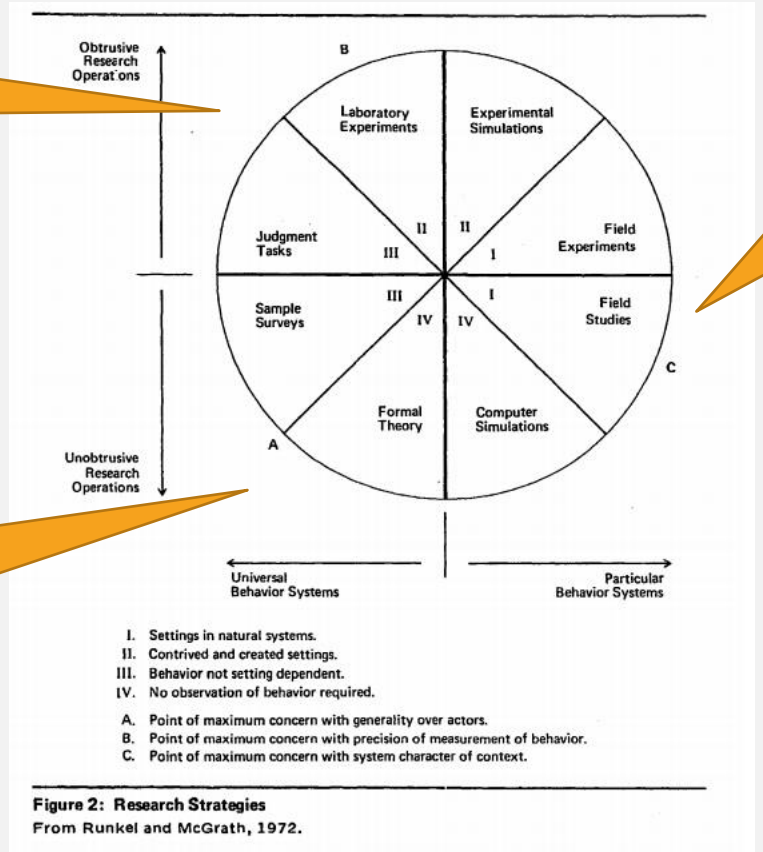


Figure 2: Research Strategies
From Runkel and McGrath, 1972.

McGRATH, JOSEPH E., Dilemmatics: "The Study of Research Choices and Dilemmas" ,American Behavioral Scientist, 25:2 (1981:Nov./Dec.) p.179

Summary

- Evaluating Interfaces with Users: Experiments
  - Lab experiments
    - Why, what, where, who, how
    - Terminologies

Post-Lecture Activity

- Read/watch these (and those in the slides)
  - Week 2 lecture slides (including the suggested readings)

- See next slide

Homework!

- Read this hypothesis:

> Deploying formative assessments in undergraduate level courses will increase students' attention on course materials during class and their final grades, relative to no formative assessments.

- Identify the following:
  - Independent variables
  - Dependent variables
  - Relationships
- How would you measure the DVs? What are the challenges you might be facing (e.g., practically, ethically)

Guest Lecture

- By PhD student Laton Vermette

- User-centered design in practice: Helping educators customize their digital classrooms
  - Introduce the Customizer platform, a recent research project conducted here at SFU that aims to give course instructors a streamlined way to customize their learning management system
  - Walk through the design process behind Customizer, including how he gathered and analyzed user feedback, iterated on a series of prototypes, and implemented a preliminary version of our design on top of Canvas
  - Describe some high-level takeaways about user-centered design and prototyping (tips, pitfalls, etc.)