Designing HCI Experiments

Chaklam Silpasuwanchai

Asian Institute of Technology

chaklam@ait.asia

March 13, 2020

Overview

Sources

- Mackenzie, Chapter 4-5, **Scientific Foundations, Designing HCI Experiments**, Human Computer Interaction: An Empirical Research Perspective, 1st ed. (2013)
- Zhao, **How to Design Controlled Experiments in HCI?** `https://www.slideshare.net/shilman/controlled-experiments-shengdong-zhao`

Reminders

- Round 2 paper reading due Friday (a hard copy in class)
  1. **Data acquisition**: use **LSL layer streaming** protocol; study how to put **markers of events**; Make sure you know which **brain regions** are of your interest, and why; use **psycopy/pygame** for making experiments and study how to **design the stimulus evoked activity**
  2. **Artifact reduction**: use **mne-python** for pre-processing/analysis; study about basic **artifact reduction techniques**: **band-pass filter, notch filter, and nyquist sampling theorem**
  3. **Dimension reduction**: study about basic reduction techniques - **PCA, ICA, common spatial patterns, wavelet transforms**
  4. **Modeling**: input your reduced components into train models - **ensembles or CNN** commonly used (you need to find out which is suitable)
  5. **Validation**: perform 5 or 10 fold **cross validation**, and other metrics
  6. **Real world use**: test with users with your trained model

Reminders

- HW11 - 13 (returned 12 and 13) (no other hws)
  1. x and y-axis should be clearly labeled
  2. Research is about **proof**; be-careful about **overclaim**; all should supported by **data**
  3. Without any statistical test, you cannot make any claim, e.g., A is faster than B even their mean of A is higher than B (due to the possibility of the incident happening by chance)
  4. You can simply state "*we found mean of A is higher than B. The SD of A and B are relatively small and thus there is chance that A is significantly higher than B. Statistical tests are needed to confirm this*"
  5. **Scatter plots** are for showing **correlations** - must plot raw data (not average) where both x and y are commonly ratio data (continuous data), **Line graphs** are for showing **trends** (does accuracy improve over time?); **bar charts** are for show **differences in groups**, i.e., nominal variables as IV and ratio data for DV; **tables** are for showing mean, SD, and p-values

What is empirical research?

- Empirical means *originating in or based on observation or experience*. It also means *capable of being verified or disproved by observation or experiment*
- Thus in HCI, empirical research is framed by **hypotheses**, where these hypotheses are verified by gathering and testing **evidence**
- In a lot of sense, empirical research covers a quantifiable, observable, reproducible aspects of interaction

Research Methods

- **Observational** Methods - include interviews, case studies, focus groups, think-aloud protocols, and so on. This approach is qualitative. As a result, this method achieve **relevance** while sacrificing precision. These methods are useful for understand the reasons underlying human behavior, as opposed to *what*, *where*, *when*
- **Experimental** Methods - also called the *scientific method* - knowledge is acquired through controlled experiments. This methodology brings **precision** while sacrificing relevance. Experimental Methods have Independent Variables and Dependent Variables
- **Correlational** Methods - looks for **relationships** between variables, e.g., privacy settings vs. IQ

Research Methods

- In HCI research, the most accepted method is **experimental method**. Of course, it is clear that experimental method will often include observational methods and correlational methods
- Golden rule is 70% quantitative (verification of effects) and 30% qualitative (tell us why)
- In experimental research, **comparative evaluation** is often done, where **proposed solution** is pit against **state-of-the art** technique as well as **baseline** technique. Baseline allows comparison of results with past studies. State-of-art allows comparison of proposed solution against the "best"
- **Methodology** involves deciding the question, the independent and dependent variables, on the people (**participants**), the hardware and software (materials or **apparatus**), the **tasks**, the **order** of tasks, the **procedure** for briefing and preparing the participants, the **data collected and analyzed**, etc.
- So what's one appropriate methodology for research in HCI? **Factorial experiments** (controlled experiments) - where participants are exposed to levels of factors while their performance is observed and measured

Comparative Evaluations

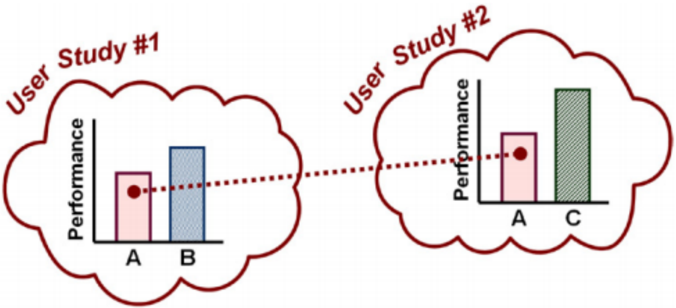

Figure: Source: Fg. 4.10 (Mackenzie)

Observe and Measure

- Observation alone is of limited value.
- There are four scales of measurement: nominal, ordinal, interval, and ratio
- We need to know these scales in order to apply appropriate data analyses
- **Nominal**: data that identify mutually exclusive categories (also known as *categorical data*, e.g., M for Male, F for Female). It is meaningless to perform mathematical manipulations on nominal data. Usually work with count or frequency (e.g., how many males do X). Nominal variables are usually our independent variables or variables used in correlational study
- **Ordinal**: provide an order or ranking. It is possible to perform comparison of ordinal data but **it is not valid to compute the mean** of ordinal data

Observe and Measure

- **Interval**: data that have **equal distances between adjacent values** but with no absolute zero, e.g., Temperature, Likert-Scale. It is meaningful to compute mean of interval data (e.g., mean temperature now is X Celsius). However, it is not meaningful to compute the **ratio** of interval data (e.g., one cannot say 20 Celsius is twice as warm as 10 Celsius)
- **Ratio**: data that have absolute zero and data can be added, subtracted, multiplied, divided, means, standard deviations. Example includes distances, words per minute, time, users' age.
- When we perform data analysis, it is important to use *standardized, normalized metric* such as **word per minute** and **error rate**, to allow easy comparisons across multiple research works

Research Questions: Example

- For our purpose, we shall set the following research question as a case study, as an example to walk-through the components of experimental research:
- How does **pie menu** - our proposed solution - compared to **linear menu** in terms of performance?

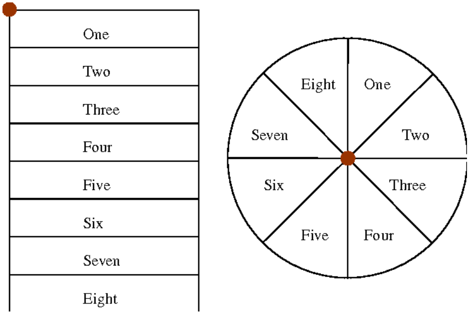

Figure: Linear menu vs. pie menu

Participants

- Choose participants (1) that are members of the **same population** of people to who results are assumed to hold and (2) make sure you have **sufficient** participants
- Way to determine how many participants: (1) see **commonality** in earlier work, (2) **power analysis** but in practice, it is rarely done because it requires knowing the variance in sample which is difficult
- In average, **12-15** participants are sufficient. One may also need to make sure the numbers of participants are sufficient such that all order of conditions are run equal
- If participants are not sufficient, the downside is that **statistical test cannot be performed** since these test relied on reliable mean and variances
- But as for usability evaluation (unlike factorial experiments), five participants are sufficient to expose 80 percent of usability problems

Reminders

- **HW14** - **16** due Friday 28 Feb: this will provide a better understanding of how to design experiments
- **First proposal draft** - INTRODUCTION AND RELATED WORK SECTION - use SIGCHI format (hard copy in class) - how long should it be?

Reminders

- **Error bars** are important when plotting bar charts as it allows us to visually inspect the variability.
- To analyze the difference between correct and incorrect one, use **bar charts**, put the correct and incorrect as x-axis, and analyze their average time and error rates for y-axis
- To analyze the difference between blocks (each block has multiple trials), use **bar charts** representing attempt 1, 2, 3 as x-axis, and analyze the average time and error rates.
- How to analyze whether number of boxes affect time and error rates, how to do?
- Use histogram to show distributions (not bar charts or line charts)
- **Trials** means one attempt, group of attempts are called **Blocks**
- **Tables** are only needed for writing average values, you almost never write the whole tables of raw data in a paper.
- Be careful what you **claim**, e.g., some of you said fatigue or boredom, which you did not measure
- Only **three questions** and not more.

Independent Variables

- Let's first decide the independent variables
- **Independent variables** (IV) are variables you **manipulate**. IV is also called a *factor*.
- In HCI, **independent variables mostly are nominal**, such as device, input method, feedback modality, selection technique, menu depth, button layout, and so on
- IV is usually manipulated across multiple **levels** (also called test conditions)
- When picking a IV, make sure you also consider different **sub-circumstances** (For text entry, one may text while *sitting*, *standing*, *walking*). In this way, your work will have two independent variables - text entry method and stance

Independent Variables

- It may be reasonable to have more than one IV, but once there are too many IV, it becomes **impossible to interpret**. For example, a design with one IV has *main effect* but no *interaction effect*. Two IV has two *main effects* and one *interaction effect*. Three IVs - there will be seven effects! **The recommended is usually two to three IVs, not too simple or too complicated. Four or above is highly not recommended.**

Independent variables	Effects					Total
	Main	2-way	3-way	4-way	5-way	
1	1	-	-	-	-	1
2	2	1	-	-	-	3
3	3	3	1	-	-	7
4	4	6	3	1	-	14
5	5	10	6	3	1	25

Figure: Source: Fg. 5.2 (Mackenzie)

Independent Variables: Example

- So what's our IV?
- Definitely, our first IV is the **menu type** which has two levels: pie menu and linear menu
- To increase our research generalizability, we can further adds more IV, for example:
  - Second IV: **menu breadth** with 3 levels: 4, 8, 12
  - Third IV: **menu depth** with 3 levels: 1, 2, 3
  - Fourth IV: **usage** with 2 levels: mobile and stationary

Thus our work is a **2 x 3 x 3 x 2 factorial design**

Dependent Variables

- Dependent variable (DV) is a **measured human behavior**
- In HCI, the most common DV is **speed** (reported in task completion time) and **accuracy** (reported in error rate)
- Others include preparation time, action time, throughput, gaze shifts, mouse-to-keyboard hand transitions, preses of BACKSPACE, target re-entries, retries, key actions, gaze shifts
- Also some creative: count of negative facial expressions, number of times users shift their gaze from on-screen keyboard to the typed text.
- When reporting, it is important to see the **common units used in earlier work**, so your work can be compared
- **Dependent variable is typically a ratio-scale** such as task-completion time, error rate, number of button clicks, scrolling events, gaze shifts, etc.

Dependent Variables: Example

- For our research question, our DV could be:
  - **Speed**: measured as completion time
  - **Accuracy**: measured as error rate
  - **Learning**: measured speed and accuracy improvements change over time

Control, Random, and Confounding Variables

- **Control** variables are factors the might influence IV such as room lighting, room temperature, background noise, selection of mouse. Researchers ought to control these variables so they are the same across during the experiment for all participants
- **Random** variables are variables that researchers may allow to vary such as age or gender of participants, personality. Usually a well-design experiment can mitigate these effects
- **Confounding** variables are condition that changes systematically with IV - using different camera for different test conditions; practice time; prior experience

Confounding Variables

(a)

ID (bits)	Amplitude (pixels) 16	32	64	128
1	*	*	*	*
2	*	*	*	*
3	*	*	*	*
4	*	*	*	*

(b)

ID (bits)	Amplitude (pixels) 16	32	64	128
1	**16**	**32**	**64**	**128**
2	**8**	**16**	**32**	**64**
3	**4**	**8**	**16**	**32**
4	**2**	**4**	**8**	**16**

(c)

W (pixels)	Amplitude (pixels) 16	32	64	128
2	*			
4	*	*		
8	*	*	*	
16	*	*	*	*
32		*	*	*
64			*	*
128				*

Figure: Source: Fg. 5.5 (Mackenzie).

- Confounding variables are sometime found in **Fitts' law experiments**
- Most Fitts law experiments use a target selection task with movement amplitude (A) and Index of difficulty (ID) as independent variables ($ID = \log_2(2A/W)$)
- If A has levels 4, 8, 16, 32cm, and ID with 1, 2, 3, 4. It will yield 16 test conditions.
- As ID increases, W decreases. Target width becomes a confounding variable. If the experiment found a main effect of ID, is the effect due to ID or to W?

Confounding Variables

- **Casual** relationship describes a *cause-and-effect* relationship between IV and DV
- Researchers have to beware for **circumstantial** relationship, that is, relationship *by circumstances or environment*

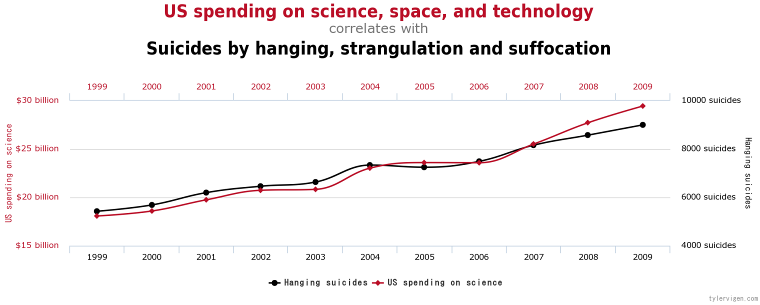

Figure: Source: Spurious Correlation by Tyler Nglen.

Experiment Validity

- Two important properties of experimental research: *internal validity* and *external validity*
- **Internal Validity** is the extent to which an effect observed is due to test conditions
  - When you are comparing two conditions, did you make sure everything else is **equal** except what you are manipulating?
  - Did you correctly **order** the experimental conditions?
  - Did you assign users to different groups in a **randomized** way?
  - Did you take care **learning effects** by applying appropriate training before the experiment or applying block design?
  - Basically, any **potential noise** lowers internal validity

Internal and External Validity

- **External Validity** is your result **generalized** across people and contexts
  - Did you choose the participants that are **representative** of the world?
  - Did you choose the experimental task that is **representative** of the world?
  - Is your system or experimental tools **representative** of the world?
  - Basically, if your result is specific to your experiment, it lowers external validity

Internal and External Validity

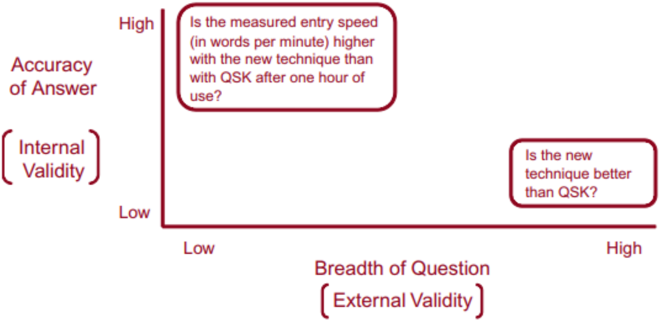

Figure: Source: Fg. 4.8 (Mackenzie)

Internal and External Validity

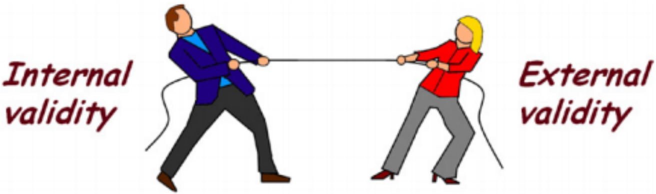

Figure: Source: Fg. 4.9 (Mackenzie)

Remote Pointing System: Validity Analysis

- Consider an experiment that compares two remote pointing devices for presentation systems
- To improve **external validity**, participants are tested in a **large room** with a **large presentation size display**, they stand, and they are positioned a few meters from the display. The other participants are engaged to act as an audience by attending and sitting around tables in the room during testing.
- But in this way, **internal validity** hurts, some participants may be **distracted** or **intimidated** by the audience. Others might have a tendency to show off, impress, or act out. Such behaviors introduce **noise**

Text-Entry: Validity Analysis

- Consider an experiment comparing two methods of text entry,
- To improve **external validity**, participants are instructed to enter whatever text they think of. The text may include punctuation symbols and uppercase and lowercase characters, and participants can edit the text and correct errors as they go
- However, **internal validity** is compromised because noise behaviors are introduced such as pondering (What should I enter next?) and fiddling with commands (How do I move the cursor back and make a correction?). Furthermore, since participants generate the text, errors are difficult to record since there is no "source text" with which to compare the entered text

Internal and External Validity

- The idea is that in research, **internal validity** cannot be compromised, i.e., researchers should put higher priority. As for external validity, researchers have to do their best in a way that their work achieve the **highest external validity possible** and also **acknowledge the limitation** in their work

Construct Validity

- is the extent to which you are **measuring things** based on what you claim
  - Measuring happiness but uses only interview or user preference or measuring text-entry performance with only speed but not errors
  - Measuring typing performance but ignore that people can type while they are walking or sitting or standing
  - Talking about habit formation but collect data using only five days experiment

Control, Random, and Confounding Variables: Example

For our case study, we can identify the following variables:

- **Control** variables are computers, experimental time, environment, instructions, etc. which shall be controlled as constant across participants
- **Random** variables are participants' age, gender, background which we cannot control, but a well-designed experiment will help
- **Confounding** variables are possible with participants' practice/learning effect. If users all try linear menu before pie menu, there is a chance pie menu achieve better than linear menu simply because of some skills acquisition. Another possible one is individual differences - a possibility where differences in performances may be due to the unbalanced groups.

Within- and between-subjects

- **Within-subjects** is when each participant is tested on each levels. Is also called *repeated measures*
- **Between-subjects** is when each participant is tested on only one level.

(a)

Participant	Test Condition		
1	A	B	C
2	A	B	C

(b)

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

Figure: Source: Fg. 5.6 (Mackenzie). a) Within-subject, b) Between subject

Within- and between-subjects

- **Within-subjects** uses **less** participants, prone to **practice effect** and thus require more **testing** (usually block design is administered). Usually preferred.
- **Between-subjects** uses **more** participants, prone to **effect of individual differences** and thus require effort to **balance** all groups. However, certain experiments require between-subject such as drug experiment or gender experiment
- **Mixed-design** uses both within-subject and between-subject in one design. For example, the experiment has two factors: block is within-subjects with perhaps 10 levels (block 1, block 2...) and handedness is between-subjects with two levels (left, right)

Within- and between-subjects: Example

- In our study, within-subject is the clear choice. Choosing between-subject will **require lots of participants** in order to balance out the effect of individual differences. The more factors (subsequently the conditions), the more participants we are required which is costly. On the other hand, within-subject is prone to **practice/learning effect** which can be easily fixed by administering **block design.**

Order Effects

- Order of conditions may affect the results, e.g., **fatigue**, **learning effects**. Thus it is necessary to *counterbalance* the order of conditions across participants
- **Latin Square** is a common method for counterbalancing. For example, if there are 12 participants and there are two test conditions, half of them does condition A first, while rest of them does condition B first.
- For more conditions, the idea is to make sure the set of order of conditions should be **equal**. Also it is important to have **sufficient participants** to fill in these sets

(a)

A	B
B	A

(b)

A	B	C
B	C	A
C	A	B

(c)

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

(d)

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

Figure: Source: Fg. 5.7 (Mackenzie).

Order Effects

- A deficiency in Latin squares of order 3 and higher is that conditions precede and follow other conditions an **unequal** number of times. In the 4 *x* 4 Latin square, for example, B follows A three times, but A follows B only once
- **Balanced Latin-square** addresses this. The top row has the sequence A, B, *n*, C, *n-1*, D, *n-2*, etc. For following rows, simply add 1
- For balanced latin-square, number of participants depend on the **number of levels**, e.g., if a factor has three levels, then the experiment requires multiple-of-3 participants (3, 6, 9, 12, etc.)

(a)

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

(b)

A	B	F	C	E	D
B	C	A	D	F	E
C	D	B	E	A	F
D	E	C	F	B	A
E	F	D	A	C	B
F	A	E	B	D	C

Figure: Source: Fg. 5.8 (Mackenzie).

Order Effects

- The one drawback of balanced Latin squares is that it only works for **even** number of test conditions
- As an example of this drawback, imagine an experiment with one IV with **three** levels (A, B, C), where each level depicts a different editing method (e.g., search and replace)
- Each participant does the task **five** times with each editing method, then does the same with other methods
- Order of levels are counterbalanced, where group 1 (participant 1, 2, 3, 4) follows **ABC**, group 2 follows **BCA**, and group 3 follows **CAB**
- This result is imbalanced because B follows A twice but A follows B only once

Order Effects

Participant	Test Condition			Group	*Mean*	*SD*
	A	B	C			
1	12.98	16.91	12.19	1	14.7	1.84
2	14.84	16.03	14.01			
3	16.74	15.15	15.19			
4	16.59	14.43	11.12			
5	18.37	13.16	10.72	2	14.6	2.46
6	15.17	13.09	12.83			
7	14.68	17.66	15.26			
8	16.01	17.04	11.14			
9	14.83	12.89	14.37	3	14.4	1.88
10	14.37	13.98	12.91			
11	14.40	19.12	11.59			
12	13.70	16.17	14.31			
Mean	15.2	15.5	13.0			
SD	1.48	2.01	1.63			

Figure: Source: Fg. 5.9 (Mackenzie).

Order Effects

- To address this imbalance for **odd** number of conditions, one may draw out all possible combinations (n!) (**full-counter balancing**) but would require more participants (here we could recruit 18 participants, each set with 3 participants). Another way is to use **Latin square** where each condition appears in each position equally - ABC, BCA, CAB

A	B	C
A	C	B
B	C	A
B	A	C
C	A	B
C	B	A

Figure: Source: Fg. 5.11 (Mackenzie).

Order Effects

- Another way to address this imbalance is to simply **randomize** the order of conditions. This is suitable when the task is **very brief**, there are many **repetitions** of the task, and there are **many test conditions**. For example, if an experiment has three factors each with multiple levels, it make sense to randomize the order
- Last, it is recommended to look at **earlier works**, to see the common acceptable counterbalancing method (usually Latin square is sufficient but again, consult with papers related to your topic)

Group Effects and Asymmetric Skill Transfer

- Counterbalancing mitigates **practice effect / learning effect**. However, this is only true when we assume that different order (e.g., A first then B and B first then A) has the same effect
- If the above assumption is incorrect, there is a **group effect** and is typically due to **asymmetric skill transfer**, i.e., there is certain skill transfer from one task to another
- Consider an IV with two levels (A:keyboard, and B:keyboard with word prediction). Participants who do A then B were found to perform better than B then A, because participants are allowed to learn the easier method first in A (Koester and Levine, 1994a)
- In this kind of case, it is recommended to use **between-subject design**

Order Effects: Example

- In our case, we have four IVs - menu type (2), breadth (3), depth (3), and usage (2)
- We can safely assume that there is no asymmetric skill transfer
- **Menu type** - full counter-balanced - is only two level so it is quite easy
- **Usage** - full counter-balanced - same as menu type
- **Breadth and width** - if we perform full counter-balance, we would need at least 144 participants (2 orders of menu types * 2 orders of usage * 6 orders of breadth * 6 orders of depth)
- What we can do is simply performing sequential orders for breadth and depth. This is doable when the difference is something of particular interest, since the differences are too obvious.Another way is to perform a randomize orders for breadth and width which is also fair.
- Thus, we would need at least 4 participants (2 orders of menu types * 2 orders of usage * 1 order of breadth * 1 order of depth). Here, it makes sense to have a multiples of 4 as our number of participants, where 16 and 20 are good numbers.

Task and Procedure

- Two objectives in designing a good task: *represent* and *discriminate*
- **Represent**: task that represent actual usage, to improve external validity
- **Discriminate**: discriminate test conditions to make sure any effects found is due to differences in test conditions, to improve internal validity
- It is nothing wrong to use **exact same task** (with slight variations like conditions) as earlier work, in fact, it is recommended within the research society to promote comparison and advancement of the field

Task and Procedure

- One specific challenge we need to tackle:
  - Users may be not performed their best when we ask them to test our technique, thus our evaluation reliability can be low
- Thus we introduce the idea of **blocks - a repeated section of an experiment that consists of multiple trials in randomized orders**.
- So how do we determine how many blocks are required?
  - More blocks are always better but should be designed in consideration of users' fatigue. Reasonable duration is 1 hour and no more than 2 hours. We also have to make sure that we have enough data points for our latter statistical analysis

Task and Procedure: Example

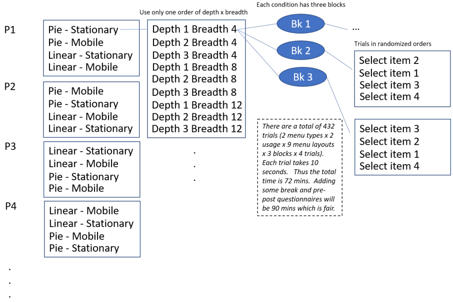

Figure: Possible experimental design

- Trial: 4 trials where each trial select certain menu item as fast and as accurate as possible
- Blocks: 3 blocks consists of multiple trials in randomized orders that repeated for each condition

Task and Procedure: Example

- Procedure:
  1. Consent form and pre-experiment questionnaires
  2. Instructions
     - First, a menu item will be shown on display to indicate target
     - Second, user presses space-bar button to indicate "start"
     - Third, user select the target menu item as fast and as accurate as possible
     - Fourth, a moment of pause before going back to first
  3. Practice trials
  4. Main experiment with breaks
  5. Post-experiment questionnaires

Questionnaire Design

- Two purposes: (1) gather information on **demographics** (age, gender, etc.) and experience with related technology, (2) gather **opinions** at the **end of experiment**

Do you use a GPS device while driving? ☐ yes ☐ no

Which browser do you use?
☐ Mozilla *Firefox* ☐ Google *Chrome*
☐ Microsoft *IE* ☐ Other (__________________)

Which browser do you use? _________________

Please indicate your age:________

Please indicate your age.
☐ < 20 ☐ 20-29 ☐ 30-39
☐ 40-49 ☐ 50-59 ☐ 60+

Questionnaire Design

- One common questionnaire to use at the end of experiment is NASA-TLX (task load index), which assesses perceived workload on six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration.

Frustration: I felt a high level of insecurity, discouragement, irritation, stress, or annoyance.

1	2	3	4	5	6	7
Strongly disagree			Neutral			Strongly agree

Questionnaire Design

- The ISO 9241-9 standard for non-keyboard input devices includes a questionnaire with 12 items to assess the comfort and fatigue experienced by participants (ISO 2000). The items are similar to those in the NASA-TLX but are generally directed to interaction with devices such as mice, joysticks, or eye trackers.

Eye fatigue:

1	2	3	4	5	6	7
Very high						Very low

Longitudinal Studies

- Sometimes, **acquisition of skills** or **improvement of performance** is of interest. In this case, testing users over a prolonged period of time called *longitudinal study* will be preferred
- In a longitudinal study, **amount of practice** is an independent variable. A typical name is *Session* with levels 1, 2, 3 and so on, where each session involves multiple trials (repetitions) of the task. Sessions composed of multiple blocks.

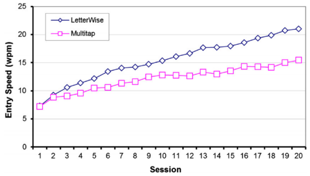

Figure: Source: Fg. 5.16 (Mackenzie). Two text-entry method were tested over 20 sessions; each session involved 30 minutes of text-entry

Longitudinal Studies

- In a longitudinal study, **crossover point** defines the point where the performance of one technique *crossover* another technique as learning progresses
- For example, in a comparison of QWERTY and Dvorak, Dvorak showed superior performance over QWERTY. QWERTY remains because the cost is too high

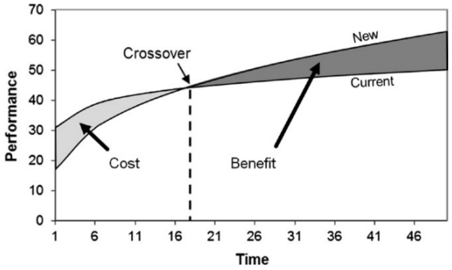

Figure: Source: Fg. 5.17 (Mackenzie).

Ethics Approval

- Since HCI is about humans, researchers must respect the safety, welfare, and dignity of human participants
- Human participants have the right to be informed of the followings: purpose of the research, methodology, any risks or benefits
- One can search for template for ethics approval

Running the Experiment

- Always useful to have a **pilot test** with one or two participants
- Important aspect of the experiment is the **instructions** given to participants; for most interaction tasks, the participant is expected to proceed quickly and accurately but comfortably
- If participants ask for **clarification**, caution must be exercised; any additional explanation that might motivate a participant to act differently from other participants is to be avoided
- Experimenter should portray himself or herself as **neutral**. Participants should not feel under pressure. Participants should also not feel too relaxed.

Reminders

- **Project Second Draft**: this Friday
- **HW20-22**: this Friday
- Most of the **latter homework** will be done together at class, to facilitate more time on project
- Install **JASP** on your PC; bring it to the classroom. We shall have three classes working on statistical analysis

Workshop

- **Problem Statement**:Which body parts are suitable for wearable vibration feedback in walking navigation for blind people
- **Independent variables**: body parts (ears, neck, wrist, hand, chest, waist, ankle, front foot, mirrored on both sides), postures (standing, normal walking, fast walking), stimulus durations (700ms, 1000ms, 1500ms, 2000ms)
- **Dependent variables**: Perceivability and subjective perferences
- **Design the rest of the experiment**, including the task and procedure, the place of experiment, the participants, the order effects, number of trials and blocks, and last, calculate the total time of the experiment

Workshop - Spoiler Answers (Don't peek!)

- This will be a design with 16 body positions x 3 postures x 4 durations x 3 trials = 576 trials
- Since each trial takes around 1s with 2.5s in between, the total time is 3.5s x 576 - 2.5s = 2013.5s / 60 = 33.558 mins - this is fair amount of time when counting time for filling questionnaires
- The **order** of body positions and stimulus duration were randomized but each body position will receive exactly 3 trials for each stimulus duration. After one posture is done, we swap to another posture. The order of posture is done using Latin-square
- The **speed of walking** must be controlled across participants (1.25m/s). The fast walking was using 4.5m/s
- **Participants** could be blind people or teenagers depending on the target audience. 15 should be nice numbers since it's the 3s multiple of the Latin-square
- **Place of environment** - could be another IV but would require another study
- After each posture, participants rated their perception of the vibration for each body position, with 1 - most difficult to perceive and 7 as easiest to perceive

Workshop (2nd round)

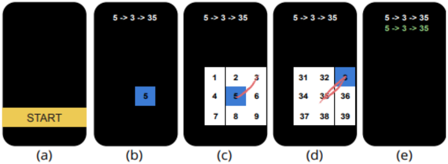

Figure: Source: Zheng et al. CHI 2018

- **Problem Statement**:We have proposed a gesture menu used in mobile phones - How does the newly proposed gesture menu compared to linear menu (baseline)?
- **Independent variables**: Input method (linear menu vs. gesture menu), Depth (1, 2, 3), Execution (guided, recall)
- **Dependent variables**: Time and error rates
- **Design the rest of the experiment**, including the task and procedure, the participants, the order effects, number of trials and blocks, and last, calculate the total time of the experiment

Workshop - Spoiler Answers

- **Depth** is one of the challenge. In D1, there are 8 possible gestures, D2 - 64 gestures, D3 - 512 gestures. To test all depths, it is possible to test completely D1 and 2 gestures, not but D3. And due to time, we definitely cannot test more than D4 and so on. For D3, we may test another 64 gestures randomizing from the sample of 512 gestures, depending on the experimental time. Since depth is an increasing complexity, the order will be strictly D1 - 2 - 3
- Another issue is the **recall** and **guided**. Obviously we should test guided before recall since there is nothing to recall.
- **Input method** can be easily fully counterbalanced
- For the **number of trials**, this needs to be prior tested before knowing how many repetitions before participants start to be good at using our menu. We found 4 trials are adequate
- This could be a design with 2 input methods x 136 gestures x 2 execution x 4 trials = 2176 trials
- Since each trial takes around 1s with 1s in between, the total time is 2s x 2176 trials - 2s = 4350s / 60 = 72.5 mins - this amount of time could be too much for participants. Thus you may want to do only 32 gestures for depth 3. Try recalculate. How much total time?

Readings For Next Week

- Mackenzie, Chapter 6, **Hypothesis Testing**, Human Computer Interaction: An Empirical Research Perspective, 1st ed. (2013)
- Yatani, Advanced Topics in Human-Computer Interaction, `http://yatani.jp/teaching/doku.php?id=2016hci:start`

Questions