Designing HCI Experiments

Chaklam Silpasuwanchai

Asian Institute of Technology

chaklam@ait.asia

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`

Research Methods

- In HCI research, the most accepted method is **experimental method**.
- **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 (1) **state-of-the art** technique and (2) **baseline** technique.
  - Baseline allows comparison of results with past studies. State-of-art allows comparison of proposed solution against the "best"

Research Question

- How does **pie menu** - our proposed solution - compared to **linear menu** in terms of performance?

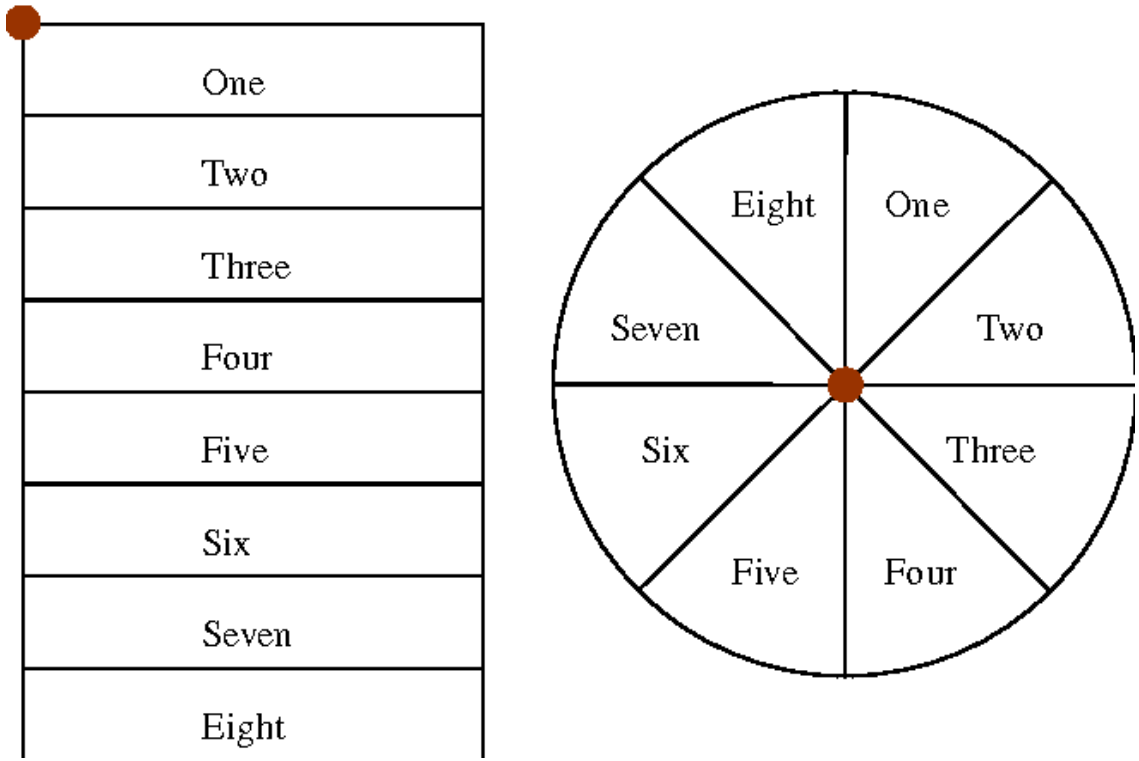

Figure: Linear menu vs. pie menu

Research Question

"Interesting" question means (1) Not obvious, (2) Can be tested, (3) Interesting to others (not only you), (4) Make sense.

Are these questions "interesting"?

- What makes people happy?
- What causes people to like Pepsi more than Coke?
- Can we predict gold prices given inflation rate?
- Is there any preference between color preferences between male and female?
- Why people watches soccer more than basketball?
- Does taller people tend to be smarter?
- Does music make us happier?
- Does playing games increase our IQ?
- Why people with similar races or religion tend not to divorce?
- Should productivity be valued more than morals?

Research Question

Science **cannot** answer all questions.

Hypothesis

Hypothesis is a statement containing some **educated guess.** Why do we need to have one? Why we just not ask the question?

A good hypothesis is (1) specific, (2) testable, (3) has clear X and y.

Are these good hypothesis?

- Will fertilizer make my lawn grow more?
- Pepsi is better than coke.
- Studying can increase exam scores.
- Heat for 1 hour cause metal to expand by 10%
- Dancing make us more happy than listening to music.

All "great" research has "great" hypothesis.

Great hypothesis comes from reading A LOT, thinking A LOT.

Hypothesis

All "great" research has "great" hypothesis.

Great hypothesis comes from reading A LOT of papers, and thinking A LOT.

Hypothesis

So what's our hypothesis?
Interesting ones could be (but not limited to)

- Pie menu is better than linear menu in general in terms of speed and accuracy
- Pie menu can be learned quickly
- Given small number of menu items, linear menu may perform equally to pie menu
- Given lots of submenus, pie menu performance over linear menu could be even more noticeable
- Similarly, pie menu performance over linear menu could be even more prominent, while walking.

Participants

Who should we pick?

Participants

Who should we pick?

- Since everyone are users, we can pick anyone. But generally, pick **target population**
- For statistical analysis, we will pick at least 12 participants. A good number is around 12-15 participants. We can also use **power analysis** or **read papers**.

Independent Variables

- IV are variables we **manipulate**. Also called **factor**. What should be our IV?

Independent Variables

- IV are variables we **manipulate**. Also called **factor**. What should be our IV?
- 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**.

Independent Variables

- Levels are sometimes called **conditions**.
- Other common IV such as feedback modality, selection technique, and so on...It is **recommended to choose between 2-3 IVs** for any experiment.
- Having too many IVs are 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!

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)

Dependent Variables

Dependent variable (DV) is **what you measure** - they **depend** on the factors. So what's our DV?

Dependent Variables

Dependent variable (DV) is **what you measure** - they **depend** on the factors. So what's our DV?

- For our case study:
  - **Speed**: measured as completion time
  - **Accuracy**: measured as error rate
  - **Learning**: measured speed and accuracy improvements change over time
- Good DVs are usually **numbers in continuous scale**
- Recommended to have **2-4 DVs**. Why not too little or too much?

Dependent Variables

- 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

Other Variables

- Other variables are **noise** variables that we want to either control (**Control** Variables), allow to vary (**Random** Variables), or do our best to mitigate (**Confounding** Variables).
- Note that a variable can be either Control, Random or Confounding, depends on how you look at them.

Control 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.
- So our study?

Control 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.
- So our study?

For our case study:

- **Control** variables for our experiment are computers, mouse, monitor, experimental time, environment, instructions, etc. which **should be** controlled as constant across participants

Random Variables

- **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
- Our study?

Random Variables

- **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
- Our study?

For our case study:

- **Random** variables are participants' age, gender, background which we cannot control, but a well-designed experiment will help. At least, we need to record these info.

Confounding Variables

- **Confounding** variables are possible noise variables that can contaminate our experiment.
- What's our possible confounding vars?

Confounding Variables

- **Confounding** variables are possible noise variables that can contaminate our experiment.
- What's our possible confounding vars?

For our case study:

- **Confounding** variables are **learning effect**, **individual differences**, and **implementation of pie menu and linear menus**

Within- and between-subjects

- Should we test all conditions with all participants?
- Or each condition with each group of participants?

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 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

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

Do you think the order of IV conditions matters?

If yes, how we should best order it?

Order Effects - Latin Square

- 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.

(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). Here A, B, C, etc. represent conditions or combined conditions.

Order Effects - Balanced Latin Square

- 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

(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 - Full Latin Square

- The one drawback of balanced Latin squares is that it only works for **even** number of test 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).

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 - Randomization

- 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**.
- Last, it is recommended to look at **earlier works**, to see the common acceptable counterbalancing method

Order Effects - Sequential

- Last way to address this imbalance is to use **sequential** order of conditions. This is suitable when the conditions you compared is of **increasing difficulty** by nature. For example, if you have two condition of a small and big width, it might be okay to always do small before big width since there is **no learning effect** anyway.
- Sequential is all about how you perceive the task whether it's an increasing difficulty task. I recommended to use this only if you are very sure.

Order Effects

How about our study?

Order Effects

- In our case, we have four IVs - menu type (MT) (2), breadth (B) (3), depth (D) (3), and usage (U) (2)
- The key here is to choose a reasonable order scheme that **minimize number of participants**, while reasonably mitigating learning effect.
- Since we have even number of conditions ($2 \times 3 \times 3 \times 2 = 36$), let's say we do **Balanced Latin Square**, we need a multiple of 36 participants, which is a lot!
- Can we further minimize participants.? Since **menu type** is our main factor, we don't want to compromise. **Usage** is only two level, so trying to do anything won't change much. **Breadth** and **Width** worth 9 conditions - since this is a lot, we can change to **Randomized** scheme or even **Sequential** scheme (since the complexity increases accordingly)
- Thus, we will have four conditions - MT1U1, MT1U2, MT2U1, and MT2U2. We will denote them as A, B, C, D. We can use balanced latin square which will give four sets: ABDC, BCAD, CDBA, DACB. Thus our number of participants will be multiples of 4; 16 and 20 are good numbers.

Task and Procedure

- It is highly recommended to use the **same task** (or with slight variations) as past work, so to promote comparison and advancement of the field. Also, they have already been well thought out.
- **Don't design your own procedure**, unless you have worked in the field for at least many years!

Task and Procedure

- What if user makes mistake?
  - use **trials**
- What if we want to monitor their learning rate
  - use **blocks** - a repeated section of an experiment that consists of multiple trials in randomized orders.
  - use **session** - which is simply composed of multiple blocks
- So how many trials and blocks? How about breaks?
  - More blocks and more breaktime are always desirable but based on **experimental time**. Why? **Reasonable duration is at max 1 hour.**

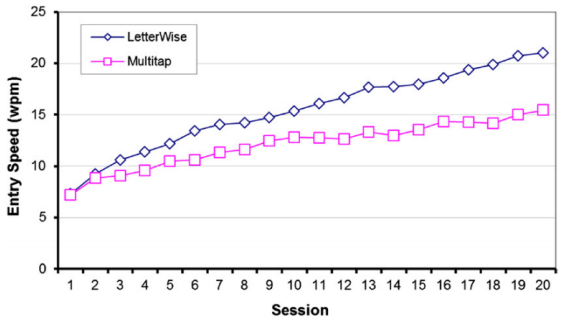

Figure: Source: Fg. 5.16 (Mackenzie). Two text-entry method were tested over 20 sessions; each session involved 30 minutes of text-entry.

Task and Procedure

For our case study:

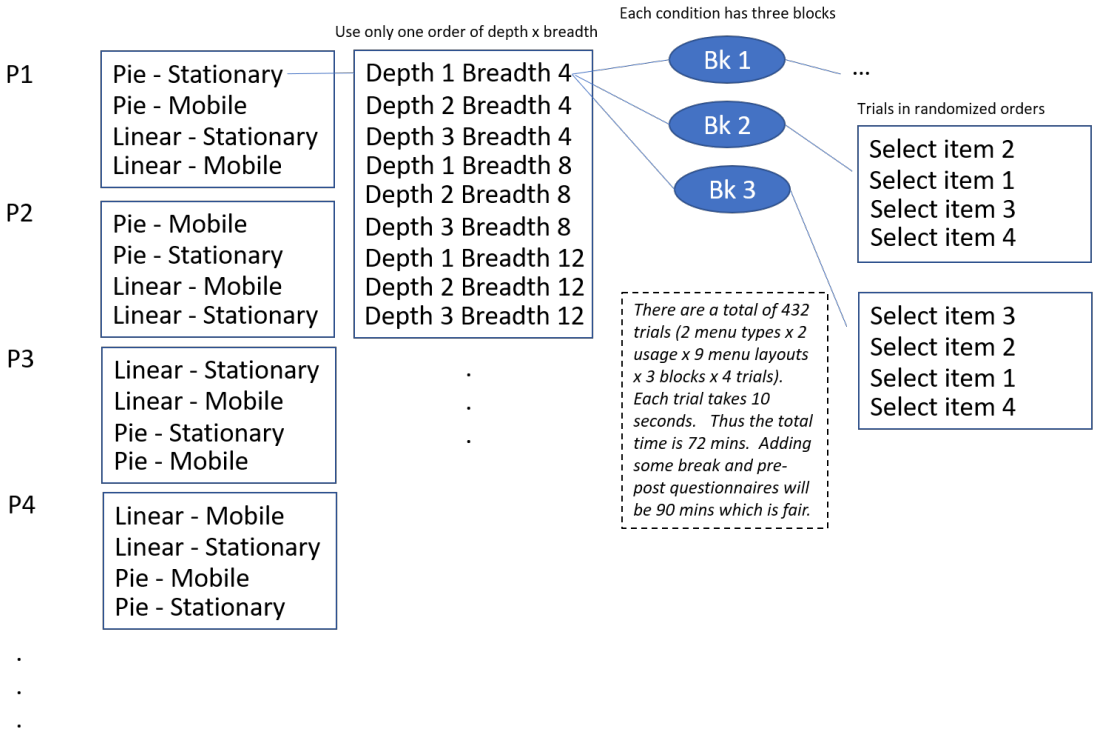

Figure: Possible experimental design

- **Trials**: 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
- **Time**: 2 **input** methods x 2 **postures** x 3 **depth** x 3 **breadth** x 3 **blocks** x 4 **trials** x 2s $=$ 864s $=$ 14.4 mins $+$ 35 **breaks** x 1 min $=$ 49.4 mins
- How did I determine the break time or blocks?

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

- **Avoid creating your own questionnaires**. Making questionnaires requires some statistical proof so it's not easy. Follow the proven ones.
- Check with your past work what questionnaires they use. **Follow them.**

Validity Analysis

- Consider an experiment that compares **two gestures technique for TV**, which experimental design?
  1. Tested in a real-world environment - **large sofa** with a **large TV**. They can watch anything. They can also eat. No instructions given.
  2. Tested in a **controlled environment** - more-controlled - task, procedure, IV, DV.

Internal and External Validity

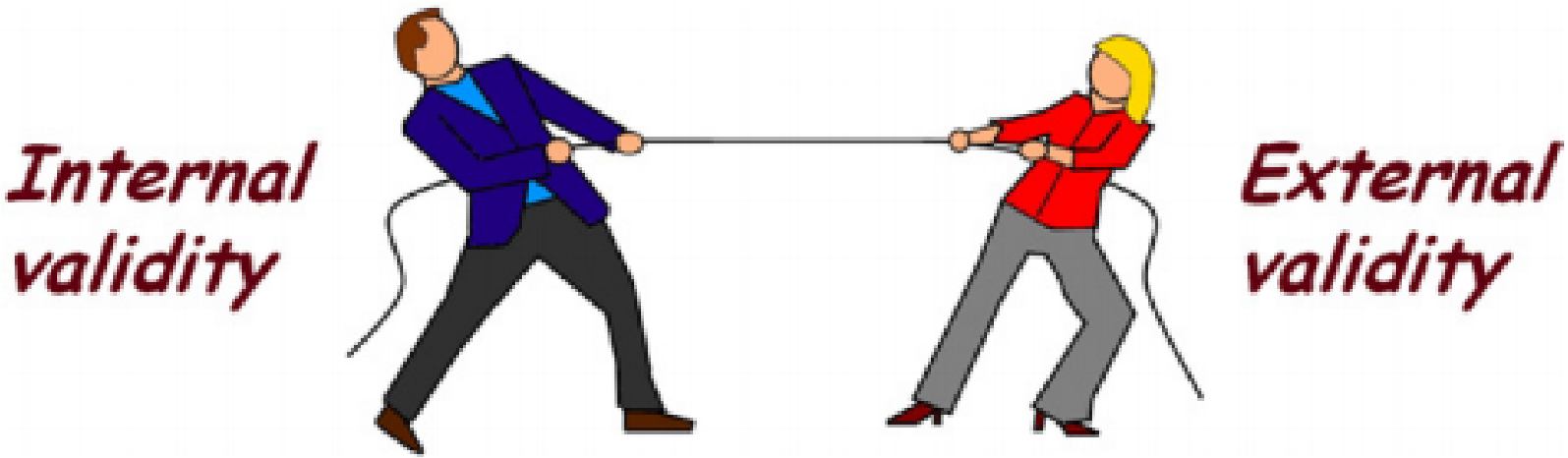

Figure: Source: Fg. 4.9 (Mackenzie)

Internal 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?

External Validity

- **External Validity** is your result **generalized** across people and contexts
  - **Representative** participant?
  - **Representative** task?
  - **Representative** tool?

Internal vs. External Validity

- The idea is that in research, **internal validity** cannot be compromised. 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
  - Measuring **typing performance** but ignore that people can type while walking
  - Talking about **habit** formation but collect data using only five days experiment

Practical suggestions

- **Reminders**: RQ → Hypothesis → Design and Implementation → Experimental Design → statistical analysis
- Always do a **pilot study.** It's almost 99% that your first experimental design will always be imperfect.
- There are **NOT only one experimental design solution; some decisions are arguable**. Of course, there are also many obviously wrong design.
- Try not to make your own task or questionnaires. **Follow papers**. This will make your work comparable, and also valid.
- **One hour** is just approximate. Some experiment is more tiring, so make sure your participants are fresh. If needed, split the experiment into several studies.
- I didn't talk much about **iterative experiment**, which is about having no clear IV, but instead iteratively explore your solution with your users directly. This is usually inefficient but only intended for people with no idea about their IV. **If you cannot think about what is your IV or hypothesis, usually because you don't understand enough.**

Link to project

- Any papers you read should be **experimental**, i.e., have clear IV and DV.
- You can attempt on any topic you want, but it should compose of experimental components, i.e., you should have clear, specific **research question**, **hypotheses**, and **experimental design**.
- Why Chaky emphasizes *experimental* papers, but does not encourage reading *exploratory* papers?

What's next

- Next coming week 3 workshops on experiment. Read and try the workshops before coming to class.

1st Workshop

Task

Given two baseline input methods: **QWERTY** and **T9**, and our proposed method: **Swiping Gestures** which we claim to be faster and more accurate. Design an experiment discussing:

- Research question
- Hypotheses
- Independent variables
- Dependent variables
- Any possible confounding/random/control variables
- Within or Between subject design
- Task
- Order
- Total experimental time

Solution

- **RQ** - How does swipe compared to QWERTY and T9 in different context, such as in different screen sizes where swipe may not have clear advantage?
- **Hypotheses**
  - H1: Swipe outperform QWERTY and T9 in terms of speed in small screen size
  - H2: In big screen size, there is no significant difference between Swipe and the other methods.
  - H3: For learning, users should be able to learn Swipe and start to show a better performance than QWERTY and T9 after some blocks of trials.
  - H4: Swipe may have lower accuracy, since it varies based on the capabilities of the word prediction.
- **IV** - **input methods** - 3 levels (QWERTY, T9, SWIPE); Let's say we have one more IV that is **Screen size** with 2 levels (Small and Large). This becomes a 3 x 2 factorial design with 6 conditions.
- **DV** - speed, accuracy, and learning over blocks

Solution

- **Control variables** - place, phone, key feedback, key font, etc.
- **Random variables** - past experience
- **Confounding variables** - implementation, task, order
- **Within-subject design** - learning can be mitigated
- **Task** - representative number of occurrences of character (perhaps use some proven dataset like `http://www.yorku.ca/mack/PhraseSets.zip`)
- **Order** - We counterbalance using balanced latin square thus we need at least a multiple of 6 participants. 12 or 18 are good number.
- **Total time** - 3 input methods x 2 screen size x 3 blocks x 6 random phrases x 28s $=$ 3024s $=$ 50 mins $+$ 5 breaks x 1 min $=$ 55 mins

2nd Workshop

Task

- **Research Question**: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 hypotheses, 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

2nd Workshop Solution

- **Hypothesis** could be something like *Upper body parts overall performed best, Longer stimulus durations may be needed for lower body parts,* Walking will generally require longer stimulus durations
- Possible **design**: 16 body positions x 3 postures x 4 durations x 3 trials = 576 trials
- Since each **trial** takes around 1s (actually 1.3s) 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
- Devices could be any arduino vibrators like Lilypads

3rd Workshop

Task

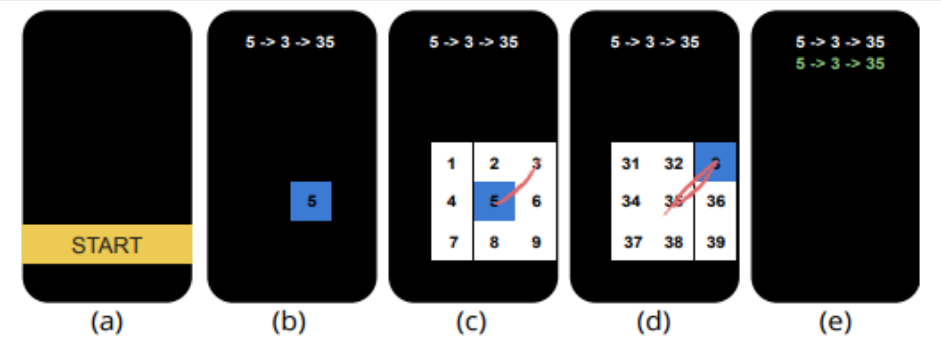

Figure: Source: Zheng et al. CHI 2018

- **Research Question**: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

3rd Workshop Solution

- **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 - 1s = 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?

4th Workshop

Task

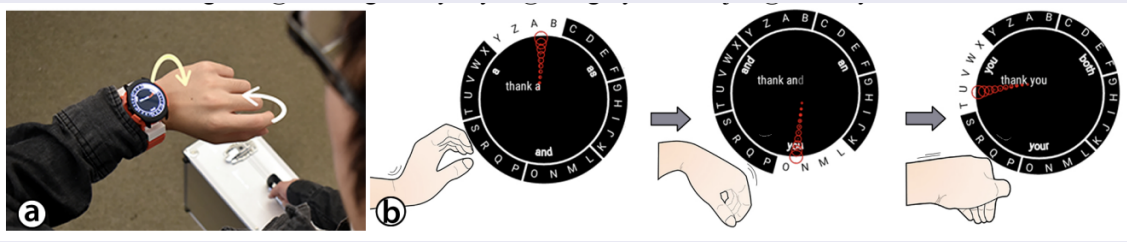

Figure: Source: Gong et al. CHI 2018; `https://dl.acm.org/doi/10.1145/3173574.3173755`, downloads: `https://cs.dartmouth.edu/~zheer/files/WrisText.pdf`

Identify (1) Research Question, (2) Hypotheses, (3) IV and DV, (4) Order Effects, (5) Task and Procedure

4th Workshop Solution

- **Research Question**: How to best design wrist-based one-handed small-form-factor text entry entry technique? Also, can users learn the technique?
- **Hypothesis**: (1) Human performance on the wrist will determine how many directions they can perform with reasonable accuracy which will then can be further optimized in the keyboard layout (study 1), (2) Users can master how they use the wrist after several days of training, and eventually can even perform eyes-free input (study 2)
- **Study 1**
  - **IV**: Target Location (8), Target Size (5)
  - **DV**: Time, Accuracy, Comfort
  - **Participants**: 15
  - **Apparatus**: Smartwatch, laptop
  - **Task**: 8 location x 5 size x 5 repetitions x 15 participants
  - **Data analysis**: ANOVA with posthoc bonferroni corrections; questionnaires for comfort ratings
  - **Findings**: A target size need to be at least 55.2 degree

Solution

- **Study 2**:
  - **IV**: Postures (hand up and down)
  - **DV**: Speed, Accuracy, Auto-completion rate (over 5 days)
  - **Participants**: 10
  - **Apparatus**: Smartwath without a ticwatch phototype; Finger-worn capacitive sensors; 27 inch screen to illustrate what input text
  - **Task**: 10 participants x 5 repetition x 18 trials x 2 postures
  - **Data analysis**: ANOVA with posthoc bonferroni corrections
  - **Findings**: User can improve through practice over time and can even perform eyes-free input with even better performance!

Questions