CSE4443 – Mobile User Interfaces

Designing A User Study

Scott MacKenzie
York University

CSE4443 – Mobile User Interfaces

Designing A User Study[1]

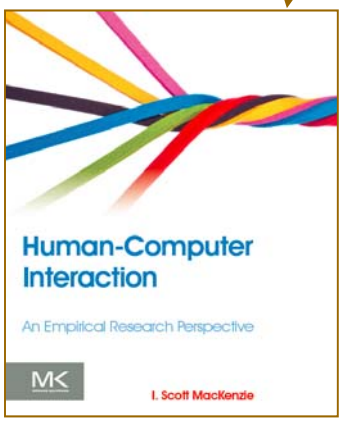
Scott MacKenzie
York University

[1] **Executive summary**

© Scott MacKenzie

Based on…

- Chapter 5 ("Designing HCI Experiments") in
- See links on CSE 4443 web page
- Free eBook access to York U students



What is a User Study?

- A "user study" is *an experiment with human participants*
- Long history in *human factors* and *experimental psychology*
- CSE 4443 → a simple user study
- The core ideas

The Goal

- Not just to evaluate a UI, but to
- Compare alternatives to determine which is better
- "Better" (like design) is a big word
- Criteria for better
  - Quantitative
    - Faster, more accurate, fewer steps, quicker to learn, etc.
  - Qualitative
    - Enjoyable, comfortable, satisfying, cool, etc.
    - Key term: *User Experience* (UX)

The Method

- *Method* → the way a user study is designed and carried out
- Methodology is critical:

> Science is method. Everything else is commentary.[1]

- What methodology?
- Don't make it up just because it seems reasonable
- Follow standards for experiments with human participants

[1] Allen Newell (cited and elaborated by Stuart Card in an invited talk at the ACM's SIGCHI conference, Austin TX, May 2012).

Getting Started

- It is difficult transitioning from the creative (ideas) to the mundane (a user study)
- Begin with…

> What are the experimental variables?

- Two variables are critical:
  - Independent variable (IV) → what you manipulate
  - Dependent variable (DV) → what you measure
- Before you can have an IV and a DV, you need a research question

Research Questions

- Typical research question:

> Can a task be performed more quickly with my new interface than with an existing interface?

- A properly formed research question identifies an IV and DV (can you spot these above?)
- IV → Interface (*new* vs. *existing*)
- DV → Speed (*more quickly*)

Causal Relationships

- A goal in doing an experiment (aka user study) is to determine a *causal relationship*
- This is possible because we balance or randomly assign conditions and participants
- In a causal relationship, changes in the DV are *caused by* the manipulations in the IV

Independent Variable

- Definition – a circumstance or characteristic that is manipulated in an experiment to elicit a change in a human response (while interacting with a computer)
- "Independent" because it does not depend on the participant (i.e., a participant cannot influence an independent variable)
- Examples:
  - interface, device, feedback mode, button layout, visual layout, age, gender, background noise, expertise, etc.
- The terms *independent variable* and *factor* are synonymous

Test Conditions

- An independent variable (IV) must have at least two levels
- The levels (aka values, settings, points of comparison) are the *test conditions*
- Name both the factor (IV) and its levels (test conditions):

Factor (IV)	Levels (test conditions)
Device	mouse, trackball, joystick
Feedback mode	audio, tactile, none
Task	pointing, dragging
Visualization	2D, 3D, animated
Search interface	Google, custom

Human Characteristics

- Human characteristics are *naturally occurring attributes*
- Examples:
  - Gender, age, height, weight, handedness, grip strength, finger width, visual acuity, personality trait, political viewpoint, first language, shoe size, etc.
- These are legitimate independent variables, but they cannot be "manipulated" in the usual sense
- Causal relationships are difficult to obtain due to unavoidable confounding variables

Dependent Variable

- A *dependent variable* is a measured human behaviour (related to interaction involving an independent variable)
- "Dependent" because it depends on what the participant does
- Examples:
  - task completion time, speed, accuracy, error rate, target re-entries, task retries, presses of backspace, expletives uttered, etc.
- Dependent variables must be clearly defined
  - Research must be reproducible!

Unique DVs

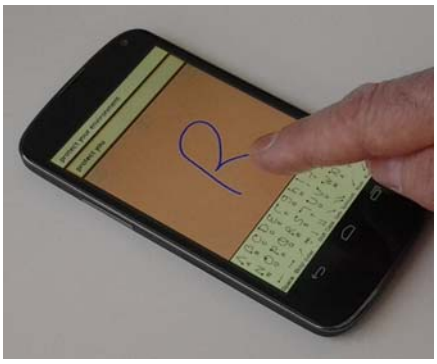
- Any observable, measurable behaviour is a legitimate dependent variable (provided it has the potential to reveal differences among the test conditions)
- So, feel free to "roll your own"
- Example: *negative facial expressions*[1]
  - Research context: user difficulty with mobile games
  - Events logged included frowns, head shaking
  - Counts used in statistically analyses, etc.
  - Clearly defined → reproducible

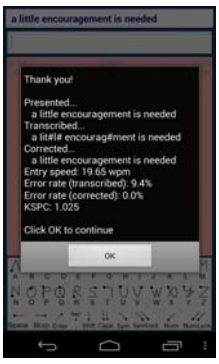
[1] Duh, H. B.-L., Chen, V. H. H., & Tan, C. B. (2008). Playing different games on different phones: An empirical study on mobile gaming. *Proceedings of MobileHCI 2008*, 391-394, New York: ACM.

Data Collection

- Obviously, the data for dependent variables must be collected in some manner
- Ideally, engage the experiment software to log timestamps, key presses, button clicks, etc.
- Planning and pilot testing important
- Ensure conditions are identified, either in the filenames or in the data columns
- Example

GraffitiExperiment





ESC

Experiment Task

- Recall the definition of an independent variable:
  - a circumstance or characteristic that is manipulated in an experiment to *elicit a change* in a human response (while interacting with a computer)
- The experiment task must "elicit a change"
- Qualities of a good task: *represent*, *discriminate*
  - Represent activities people typically do
    - Improves external validity (ability to generalize)
  - Discriminate among the test conditions
    - Improves internal validity (finding differences that are real)

Task Examples

- Usually the task is self-evident
- Research idea → new widgets for creating entry in calendar app
  - Experiment task → create entry in calendar app using (a) new widgets and (b) conventional method
- Research idea → auditory feedback for programming GPS destination
  - Experiment task → program destination into GPS device using (a) musical sounds (b) natural sounds (c) conventional method

Procedure

- The *procedure* encompasses everything that occurs with participants
- The procedure includes the task (obviously), but everything else as well…
  - Arriving, welcoming
  - Signing a consent form
  - Instructions given to participants about the experiment task (next slide)
  - Demonstration trials, practice trials
  - Rest breaks
  - Administering of a questionnaire or an interview

Instructions

- Very important (best to prepare in advance; write out)
- Often the goal in the experiment task is "to proceed as quickly and accurately as possible but at a pace that is comfortable"
- Other instructions are fine, as per the goal of the experiment or the nature of the tasks, but…
- Give the same instructions to all participants
- If a participant asks for clarification, do not change the instructions in a way that may cause the participant to behave differently from the other participants

Participants

- Researchers want experimental results to apply to people not actually tested – a population
- Population examples:
  - Computer-literate adults, teenagers, children, people with certain disabilities, left-handed people, engineers, musicians, etc.
- For results to apply generally to a population, the participants tested must be…
  - Members of the desired population
  - Selected at random from the population
- True random sampling is rarely done (consider the number and location of people in the population examples above)
- Some form of *convenience sampling* is typical

How Many Participants?

- Too few → experimental effects fail to achieve statistical significance
- Too many → statistical significance for effects of no practical value
- The correct number… (drum roll please)
  - Use the same number of participants as used in similar research[1]
- 4443 project → 8 minimum

[1] Martin, D. W. (2004). *Doing psychology experiments* (6th ed.). Pacific Grove, CA. Belmont, CA: Wadsworth.

Questionnaires

- Questionnaires are given in most user studies
- Two purposes
  1. Collect information about the participants
     - Demographics (gender, age, first language, handedness, visual acuity, etc.)
     - Prior experience with interfaces or interaction techniques related to the research
  2. Solicit feedback, comments, impressions, suggestions, etc., about participants' use of the experimental apparatus
- Questionnaires, as an adjunct to a user study, are usually brief

Information Questions

- Questions constructed according to how the information will be used

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? _____

Frustration: My level of insecurity, discouragement, irritation, stress, or annoyance was

1	2	3	4	5	6	7
Very low						Very high

Participant Feedback

- Using NASA Task Load Index (TLX):

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

- ISO 9241-9:

Eye fatigue:

1	2	3	4	5	6	7
Very high						Very low

Within-subjects, Between-subjects

- Two ways to assign conditions to participants:
  - *Within-subjects* → each participant is tested on each condition (aka *repeated measures*)
  - *Between-subjects* → each participant is tested on one condition only
  - Example: An IV with three test conditions (A, B, C):

Within-subjects

Participant	Test Condition		
1	A	B	C
2	A	B	C

Between-subjects

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

Within-subjects, Between-subjects (2)

- Within-subjects advantages
  - Fewer participants (easier to recruit, schedule, etc.)
  - Less "variation due to participants"
  - No need to balance groups (because there is only one group!)
- Within-subjects disadvantage
  - Order effects (i.e., interference between conditions)
- Between-subjects advantage
  - No order effects (i.e., no interference between conditions)
- Between-subjects disadvantage
  - More participants (harder to recruit, schedule, etc.)
  - More "variation due to participants"
  - Need to balance groups (to ensure they are more or less the same)

Within-subjects, Between-subjects (3)

- Sometimes…
  - A factor must be assigned within-subjects
    - Examples: Block, session (if learning is the IV)
  - A factor must be assigned between-subjects
    - Examples: gender, handedness
  - There is a choice
    - In this case, the balance tips to within-subjects (see previous slide)
- With two factors, there are three possibilities:
  - both factors within-subjects
  - both factors between-subjects
  - one factor within-subjects + one factor between-subjects (this is a *mixed design*)

Order Effects, Counterbalancing

- Only relevant for within-subjects factors
- The issue: *order effects* (aka *learning effects*, *practice effects*, *fatigue effects*, *sequence effects*)
- Order effects offset by *counterbalancing*:
  - Participants divided into groups
  - Test conditions are administered in a different order to each group
  - Order of administering test conditions uses a Latin square
  - Distinguishing property of a Latin square → each condition occurs precisely once in each row and column (next slide)

Latin Squares

2 x 2

A	B
B	A

3 x 3

A	B	C
B	C	A
C	A	B

4 x 4

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

5 x 5

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

Balanced Latin Square

- With a balanced Latin square, each condition precedes and follows each other condition an equal number of times
- Only possible for even-orders
- Top row pattern: A, B, *n*, C, *n* – 1, D, *n* – 2, …

4 x 4

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

6 x 6

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

Example

- User study to determine if three soft keyboards (A, B, C) differ in the amount of time to do a common editing task:

  Replace one 5-letter word with another, starting one line away.

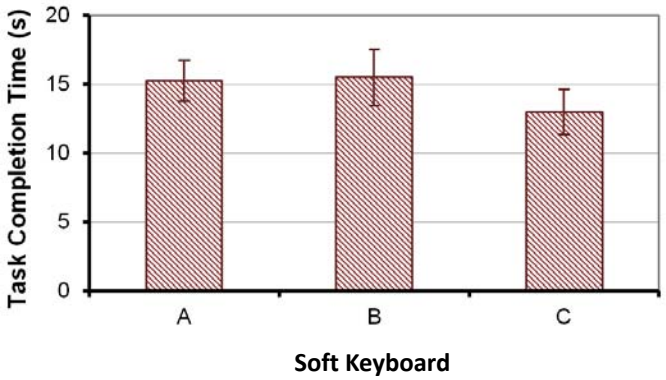
- Conditions are assigned within-subjects
- Twelve participants are recruited and divided into three groups (4 participants/group)
- Methods administered using a 3 × 3 Latin Square (2 slides back)
- Results (next slide)

Results - Data

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			

Group effect is small
∴ Counterbalancing worked!

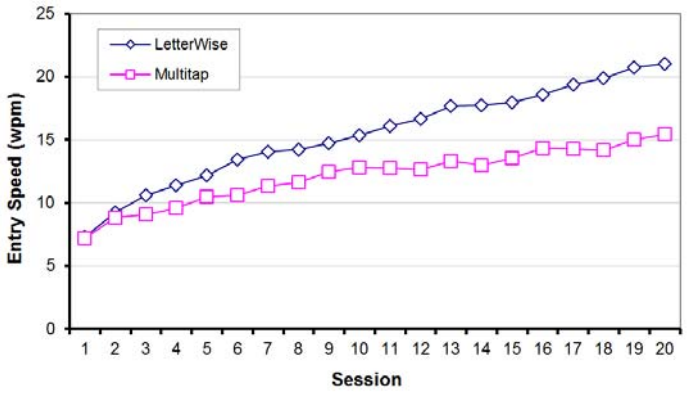
Results - Chart



Longitudinal Studies

- Sometimes instead of "balancing out" learning effects, the research seeks to study learning
- If so, a *longitudinal study* is conducted
- "Practice" is the IV
- Participants are practiced over a prolonged period of time
- Practice units: blocks, sessions, hours, days, etc.
- Example on next slide

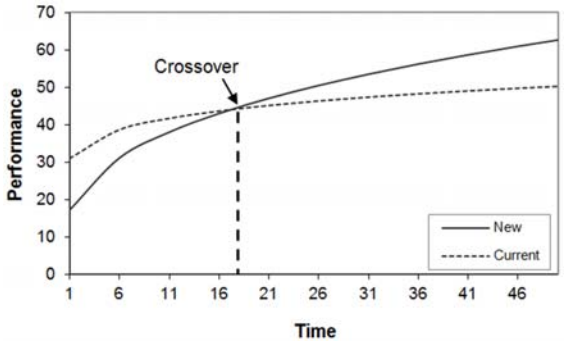
Longitudinal Study – Results[1]



[1] MacKenzie, I. S., Kober, H., Smith, D., Jones, T., & Skepner, E. (2001). LetterWise: Prefix-based disambiguation for mobile text entry. *Proceedings of the ACM Symposium on User Interface Software and Technology - UIST 2001*, 111-120, New York: ACM.

The New vs. The Old

- Sometimes a new technique will initially perform poorly in comparison to an established technique
- A longitudinal study will determine if a crossover point occurs and, if so, after how much practice (see below)



Thank You