SWEN 422 Lecture 6
Evaluation 2

Dr Jennifer Ferreira

15 March 2024

Agenda

- Review of previous lecture
- Things to consider when doing evaluations
- How many participants?
- Participants' rights and getting consent
- User Feedback
- When not to evaluate?

Summative Research in HCI

- Understand **behaviour** with the new system
  - Is the web-based tool for supporting holistic building energy management usable?
- Understand **attitudes** to the new system
  - Will clinicians adopt the “CanRisk” tool for predicting risk of breast and ovarian cancer?
- Evaluate performance of a working **prototype**
  - Usability Assessments of STAR-Vote
- Based on our findings we may want to
  - Refine requirements for the new system
  - Establish that the new system is “better” / “usable”
  - Establish that the new system is “ready” for release / “fit-for purpose”
  - Establish that users will accept & use the new system
  - Conduct further studies

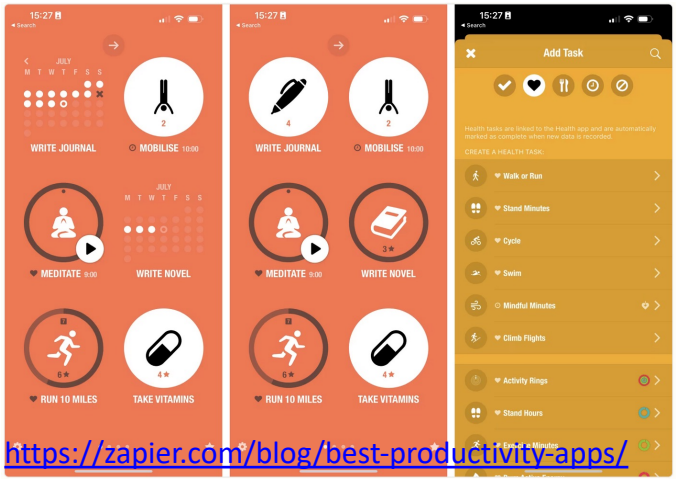
Measuring usability

- Measuring ***learnability***
- Measuring ***effectiveness/ efficiency***
- Measuring ***memorability/recall***
- Measuring user ***satisfaction***
- Measuring ***errors***

Things to consider when conducting evaluations

- **Biases** *(Qual & Quant)*: Are there biases that influence the results?
- **Reliability** (*Quant*): Does the method produce the same results on separate occasions?
- **Validity** (*Quant*): Does the method measure what it is intended to measure?
- **Ecological validity** *(Quant)***/Transferability** *(Qual)*: Does the environment of the evaluation distort the results?
- **Credibility** (*Qual*): Can others recognise the experiences contained within the study?
- **Dependability** (*Qual*): Can another researcher follow the decision trail?

Quantitative study example



https://zapier.com/blog/best-productivity-apps/

Research Question: Does the introduction of a new task management mobile application improve users' productivity compared to existing methods?

Measure: Number of tasks completed within a specified time period (M1).

Reliable?

Some users report their own M1 while some users are measured by a researcher in a lab

Valid?

Logger that starts timing from when the user opens the app until they move away from the app.

Qualitative study example

Research Question: How do medical staff collaborate in an emergency department?
Observations: In the ED.

Credible?
Medical staff: “Do we do this?”

Dependable?
Short interviews in a university meeting room.
Lab based usability study of ventilator.

Task selection bias

- We create tasks for users to complete -> it is possible to complete
- The user knows this

Social desirability bias

- Users tell you what they think you want to hear/what they think will make them look good

Confirmation bias

- When a tester pushes users into giving the feedback that confirms their own assumptions or feelings about the software

How many participants?

- Are we attempting to **improve a system** or to find **universal knowledge about** people or systems?

- Are we looking for **all** usability problems or the **major usability** problems?

"When doing any kind of user research, you can study **large numbers shallowly** or **small numbers in depth**..."(Gilmore, D. Understanding and overcoming resistance to ethnographic design research. Interactions, 9, 3 (2002), 29-35)

How many participants?

"The **validity** of a usability issue depends **not on the number** of participants who exhibited the issue, but rather the ability of the usability professional to create a **plausible and rational account** of the exhibited behaviour."

(Katz & Rohrer, 2004)

How many participants?

- Sample size is relevant for **statistical significance** (and confidence intervals).
  - Levels of confidence about one thing being "better" than another
  - Levels of confidence that the results are not based on chance

Why Care about Statistical Significance? By William Hudson

Deriving a Problem Discovery Sample Size by Jeff Sauro

Calculator

Participants' rights and getting consent

- Participants need to be told why the evaluation is being done, what they will be asked to do and informed about their rights.
- **Information sheets and consent forms** provide this information and act as a contract between participants and researchers.
- The evaluation needs to be approved by an ethics board (HEC).

https://en.wikipedia.org/wiki/Hawthorne_effect

Hawthorne Effect

- People change their behaviour when they know they are being observed
- Not the name of a person but a *factory*
- Based on studies of the effects on light levels on worker productivity
- Result: productivity increased when light levels changed / when any variable changed

Covert/undercover usability testing

- Covert naturalistic observation
  - observing behaviours in their natural contexts without any intervention or influence by the researcher and without participants knowing that they're being observed.
  - Pro: No Hawthorne Effect
  - Con: Can't ask questions, keep your distance, ethics
  - Popular in psychology, anthropology, and other social sciences

Covert/undercover usability testing

- A/B testing
  - Two versions are tested (online)
  - Randomly assign users to 2 groups
  - Most basic form of randomised controlled experiment
  - “Control: 15% (+/- 2.1%), Variation: 18% (+/- 2.3%).”
- “Do not know” is different to deception
  - E.g. participants complete a quiz, and are falsely told that they did very poorly, regardless of their actual performance

Valuing the user's design feedback

- 2 types of data from usability evaluations:
  - *interaction data* – screen recordings, system logs, notes from think-aloud protocols made by the researcher
  - *design feedback* – user's reflections and comments, e.g. "I think this button should be larger."

Valuing the user's design feedback

- 2 types of data from usability evaluations:
  - *interaction data* – screen recordings, system logs, notes from think-aloud protocols made by the researcher
  - *design feedback* – user's reflections and comments , e.g. "I think this button should be larger."

What to do with this?

Valuing the user's design feedback

- 2 types of data from usability evaluations:
  - *interaction data* – screen recordings, system logs, notes from think-aloud protocols made by the researcher
  - *design feedback* – user's reflections and comments , e.g. "I think this button should be larger."

What to do with this?

- Specialised contexts of use
- Emergency responders, airline pilots, neurosurgeons
- Participatory design
- Participatory Design in Practice

Valuing the user's design feedback

- 2 types of data from usability evaluations:
  - *interaction data* – screen recordings, system logs, notes from think-aloud protocols made by the researcher
  - *design feedback* – user's reflections and comments , e.g. "I think this button should be larger."

X

Jakob Nielsen: https://www.nngroup.com/people-jakob-nielsen-photos/

- "Don't listen to users"
- Biased
- Imperfect memory
- Rationalise

In which situations can we drop usability evaluation?

What's the motivation here?

- Doing less work?
- Not having to deal with users?
- Not having to deal with ethics?
- Not wanting to deal with critique?

What's the motivation here?

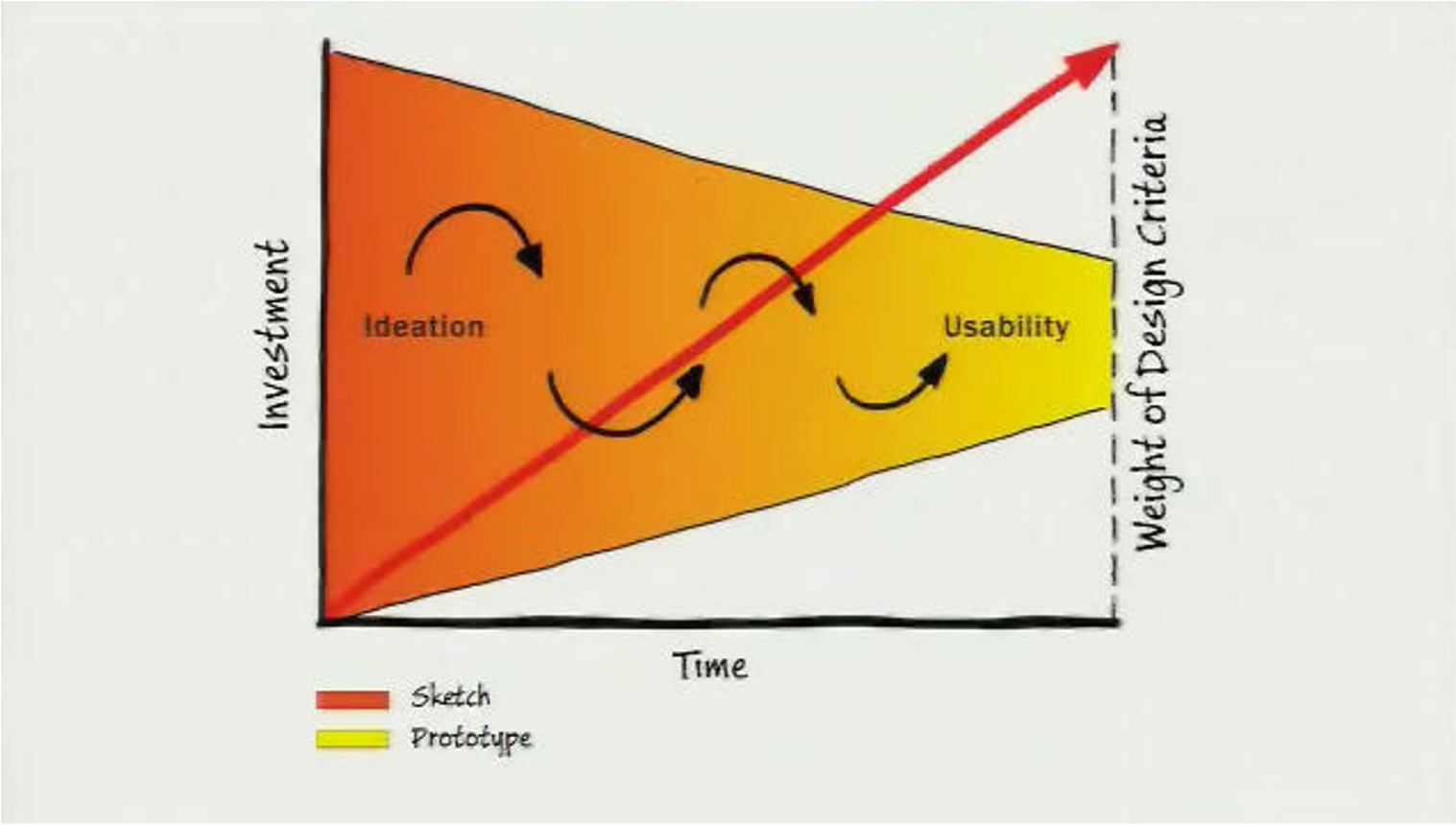
- ~~Doing less work?~~
- ~~Not having to deal with users?~~
- ~~Not having to deal with ethics?~~
- ~~Not wanting to deal with critique?~~
- Selling like hot cakes
- Backend system that doesn't interface with users (maybe)
- Greenberg & Buxton:
  - Stifles early design ideas -> getting the right design vs. getting the design right
  - Measuring the "measurables" (duration, clicks, etc.) is not helping

Further reading

- Følstad, A. Users' design feedback in usability evaluation: a literature review. *Hum. Cent. Comput. Inf. Sci.* **7,** 19 (2017). https://doi.org/10.1186/s13673-017-0100-y
- Greenberg, S., & Buxton, B. (2008, April). Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 111-120).
- Jakob Nielsen: How Many Test Users in a Usability Study?
- Gilmore, D. Understanding and overcoming resistance to ethnographic design research. Interactions, 9, 3 (2002), 29-35.
- Chapter 14 in Preece, Jenny, et al. *INTERACTION DESIGN : BEYOND HUMAN-COMPUTER INTERACTION*, Wiley, 2015. *ProQuest Ebook Central*, http://ebookcentral.proquest.com/lib/vuw/detail.action?docID=4901891
- Research involving Deception: https://research.oregonstate.edu/irb/research-involving-deception

Issues in Evaluation

- How many users enough?
  - https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.498.9565&rep=rep1&type=pdf
- Open science in HCI
- Adaptations required: https://link-springer-com.helicon.vuw.ac.nz/article/10.1007/s42979-020-00424-4, https://dl-acm-org.helicon.vuw.ac.nz/doi/10.1145/3409118.3475135
- https://learning.oreilly.com/library/view/quantifying-the-user/9780123849687/xhtml/CHP008.html#ST0025



Design “Funnel” - [Buxton]

Open science in HCI

- https://dl-acm-org.helicon.vuw.ac.nz/doi/10.1145/3490554
- https://www-degruyter-com.helicon.vuw.ac.nz/document/doi/10.1515/pik-2015-0009/html

Models of interaction

- Rigorous notation/description
- GOMS, BNF notation, TAG, etc. (https://www.cs.ox.ac.uk/files/3423/PRG97.pdf)
- Id-book ch 8