Usability Evaluation Methods I

SPRING 2020

DR MARK FREEMAN

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Week	Topic	Reading
1	Introduction to HCI; Design Principles	Chapter 1
2	User-Centred Design Process	Chapter 2 & 3 / Gould et al. (1987)
3	User Interaction and Interfaces	Chapters 4, 5 & 6 / Shin et al. (2017)
4	Interaction Design and Development I	Chapters 7 & 8
5	Interaction Design and Development II	Chapters 9 & 10
6	Interaction Design and Development III	Chapters 11 & 12
7	Information Presentation and Design Patterns	
8	Usability Evaluation Methods I	Chapters 13 / Borsci et al. (2015)
9	Usability Evaluation Methods II	Chapter 14, 15 & 16
10	Accessibility and Special Issues in HCI	Online: WCAG2.0
11	Models, Theories and Risks	MacKenzie (1992)
12	Mixed Reality and Future HCI	
13	Subject Revision	

This Week

- System Testing
- Introduction to UEMs
  - Usability Evaluation Methods
- User-Based Methods

Subject Description

- The subject provides students with an understanding of Human Computer Interaction (HCI) principles and practices, and how to apply them in the context of developing usable interactive computer applications and systems. The subject also emphasises the importance of taking into account contextual, organisational, and social factors in the design of computer systems. Students will be taken through the analysis, design, development, and evaluation of user interfaces. They will acquire hands-on design skills through an interaction design project. The subject will cover topics including user-centred design, the development process, prototyping, usability testing, measuring and evaluating the user experience and accessibility.

Subject Learning Outcomes (SLOs)

- On successful completion of this subject, students will be able to:
  1. Identify and describe HCI principles and design issues.
  2. Discuss and justify HCI solutions based on design principles.
  3. Demonstrate an understanding of the HCI design process.
  4. Acquire skills to design and implement user-centred design.
  5. Select and use suitable methods of measuring and evaluating the user experience.

After a system has been developed…

THIS WEEK WE NEED TO SHIFT OUR THINKING

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

System Testing

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Testing Concepts

- **Testing** – the process of examining a component, subsystem, or system to determine its operational characteristics and whether it contains any defects
- **Test case** – a formal description of a starting state, one or more events to which the software must respond, and the expected response or ending state
  - Defined based on well understood functional and non-functional requirements
  - Must test all normal and exception situations
- **Test data** – a set of starting states and events used to test a module, group of modules, or entire system
  - The data that will be used for a test case
- **Unit test** – tests of an individual method, class, or component before it is integrated with other software

Integration Testing

- **Integration test** – tests of the behavior of a group of methods, classes, or components
  - Interface incompatibility— For example, one method passes a parameter of the wrong data type to another method
  - Parameter values— A method is passed or returns a value that was unexpected, such as a negative number for a price.
  - Run-time exceptions— A method generates an error, such as “out of memory” or “file already in use,” due to conflicting resource needs
  - Unexpected state interactions— The states of two or more objects interact to cause complex failures, as when an class method operates correctly for all possible Customer object states except one

System, Performance, and Stress Testing

- **System test** – an integration test of an entire system or independent subsystem
  - Can be performed at the end of each iteration
  - Can be performed more frequently
- **Performance test or stress test** – an integration and usability test that determines whether a system or subsystem can meet time-based performance criteria
  - Response time – the desired or maximum allowable time limit for software response to a query or update
  - Throughput – the desired or minimum number of queries and transactions that must be processed per minute or hour

User Acceptance Testing

- User acceptance test – a system test performed to determine whether the system fulfills **user requirements**
- May be performed near the end of the project (or at end of later project iterations)
- A very formal activity in most development projects

- Payments can be tied to passing tests

Usability

FROM WEEK 1

- **Usability**: “The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.” (ISO 9241-11)
- “**User experience** (UX) is an approach to product development that incorporates direct user feedback throughout the development cycle (human-centered design) in order to reduce costs and create products and tools that meet user needs and have a high level of usability (are easy to use).” (UXPA 2013)

Usability Evaluation Methods (UEMs)

PREECE ET AL. (2019) – TEXT CATEGORIES

Quick and Dirty Evaluations

- Informal (throughout UCD)

Usability Testing

- Measuring user performance

Field Studies

- In natural settings

Predictive Evaluation

- Expertise

Usability Evaluation Methods (UEMs)

- User-Based Evaluation
  - Performance, non-verbal behaviour, attitude, cognition, stress and motivation
  - user experience and the interconnectivity that exists between usability and user experience (Vermeeren et al., 2010)
    - *(could include: Quick and Dirty, Usability Testing, Field Studies)*
- Expert-Based Evaluation
  - Conformance and attitude
    - *(could include: Quick and Dirty, Predictive Evaluation)*
- Theory Based Evaluation
  - Performance (idealised)
    - *(could include: Predictive Evaluation)*

(Sweeney et al. 1993)

Evaluation methods

Method	Controlled settings	Natural settings	Without users
Observing	X	X	
Asking users	X	X	
Asking experts		X	X
Testing	X	x*	
Modeling			X

* Some testing can be performed but with greater difficulty

errāre hūmānum est
(to err is human)

WHY CONDUCT TESTING?

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Human error

- “Error will be taken as a generic term to encompass all those occasions in which a planned sequence of mental or physical activities fails to achieve its intended outcome, and when these failures cannot be attributed to the intervention of some chance agency.” (Reason, 1990)

- Slip
  - Understand system and goal
  - Correct formulation of action
  - Incorrect action
- Mistake
  - May not even have right goal

- How can we fix?
  - Slip --- better interface design
  - Mistake --- better user understanding of system

Consider, Task Analysis

- Task Analysis is used to identify the sequence of tasks required to complete an activity
- It involves breaking down tasks into discrete steps, and noting the order in which they occur
- It should also expose areas of under productivity which a computerised system could automate
- Any of the following methods can be used to collect information about a system:
  - Observation (ethnography)
  - Interviews with users
  - User logging

From Task Analysis

- Who is going to use system?
- What tasks do they now perform?
- What tasks are desired?
- How are the tasks learned?
- Where are the tasks performed?
- What's the relationship between user & data?
- What other tools does the user have?
- How do users communicate with each other?
- How often are the tasks performed?
- What are the time constraints on the tasks?
- What happens when things go wrong?

How can these then be evaluated?

Will help with gathering the requirements for testing

Why, what, where and when to evaluate

- Iterative design & evaluation is a continuous process that examines:

- **Why**: to check users' requirements and that they can use the product and they like it.
- **What**: a conceptual model, early prototypes of a new system and later, more complete prototypes.
- **Where**: in natural and/or laboratory settings.
- **When**: throughout design; finished products can be evaluated to collect information to inform new products.

Types of User Evaluations (Environment)

- **Controlled settings** involving users
  - Usability Testing
  - Usability Experiments
    - In laboratories and living labs.

- **Natural settings** involving users
  - Field studies
  - In the wild studies
    - to see how the product is used in the real world.

- Testing can be conducted remotely

Usability testing & research

Usability testing

- Improve products
- Few participants
- Results inform design
- Usually not completely replicable
- Conditions controlled as much as possible
- Procedure planned
- Results reported to developers

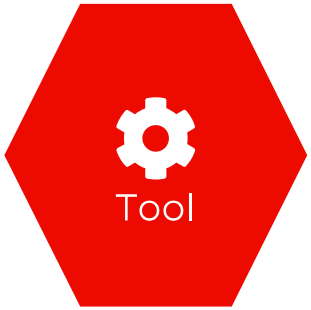
Testing conditions – Lab Environments

- Usability lab or other controlled space
- Emphasis on:
  - selecting representative users;
  - developing representative tasks.

- 5-10 users typically selected
- Tasks usually around 30 minutes
- Test conditions are the same for every participant
- Informed consent form explains procedures and deals with ethical issues

An approach…Involving Users

Step 1: **R**equirements	• Determine the goals of the user task-based testing • Identify the (types of) end users to be evaluated
Step 2: **E**nvironment	• Determine which contexts-of-use will be evaluated • Design the environment for the task-based testing (simulated (laboratory) versus actual)
Step 3: **S**etup	• Select the equipment (mobile devices, POV camera(s) & software) • Develop the scenarios
Step 4: **T**esting	• Conduct and record the scenarios with real users, using real devices in the designed environment
Step 5: **E**valuation	• Evaluate scenarios at the level of individual operations • Identify user errors, and model user experience results
Step 6: **R**eporting	• Report to developers and/or management on the task-based testing • Modify development guidelines with reference to testing results



Controlled Setting – UQ Lab

Usability testing & field studies can compliment

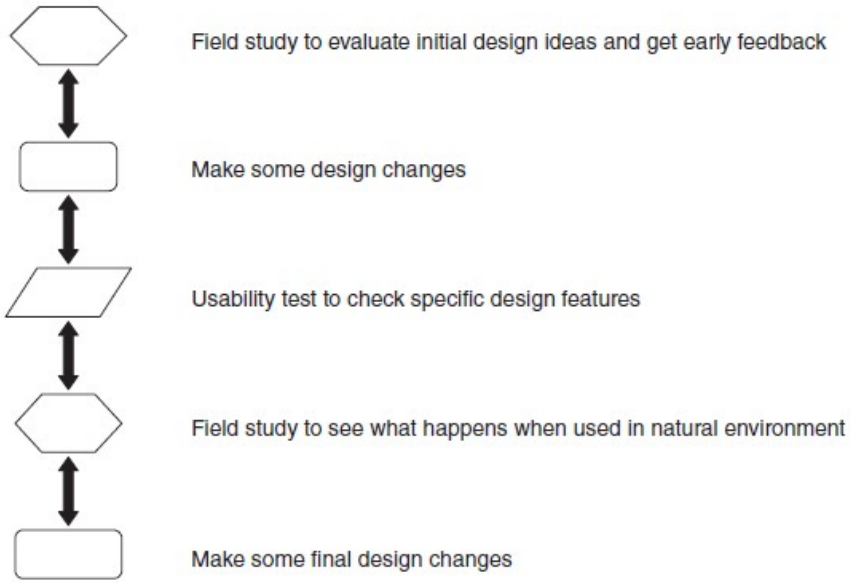
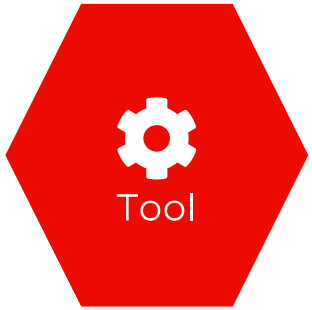


Figure 13.1 Example of the way laboratory-based usability testing and field studies can complement each other



Participants' rights and getting their consent

- Participants need to be told why the evaluation is being done, what they will be asked to do and their rights.
- Informed consent forms provide this information.
- The design of the informed consent form, the evaluation process, data analysis and data storage methods are typically approved by a high authority, e.g. Institutional Review Board.

At UOW: http://www.uow.edu.au/research/ethics/human/index.html

Usability Testing

- Involves recording performance of typical users doing typical tasks.
- Controlled settings.
  - Users are observed and timed.
  - Data is recorded on video & key presses are logged.
  - The data is used to calculate performance times, and to identify & explain errors.
  - User satisfaction is evaluated using questionnaires & interviews.
  - Field observations may be used to provide contextual understanding.

Experiments & Usability Testing

- Experiments test hypotheses to discover new knowledge by investigating the relationship between two or more variables.
- Usability testing is applied experimentation.
- Developers check that the system is usable by the intended user population for their tasks.

Usability Testing

- Goals & questions focus on how well users perform tasks with the product.
- Comparison of products or prototypes is common.
- Focus is on time to complete task & number & type of errors.
- Data collected by video & interaction logging.
- Testing is central.
- User satisfaction questionnaires & interviews provide data about users' opinions.

Experiments

Test hypothesis

Predict the relationship between two or more variables.

Independent variable is manipulated by the researcher.

Dependent variable influenced by the independent variable.

Typical experimental designs have one or two independent variables.

Validated statistically & replicable.

Participant design

Design	Advantages	Disadvantages
Different	No order effects	Many subjects & individual differences a problem
Same	Few individuals, no individual differences	Counter-balancing needed because of ordering effects
Matched	Same as different participants but individual differences reduced	Cannot be sure of perfect matching on all differences

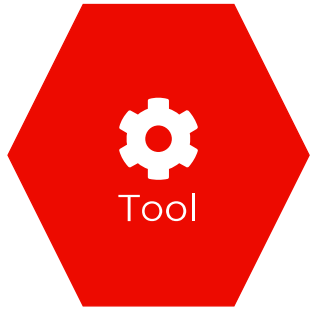
Experimental designs

- Different participants - single group of participants is allocated randomly to the experimental conditions.
- Same participants - all participants appear in both conditions.
- Matched participants - participants are matched in pairs, e.g., based on expertise, gender, etc.

Types of data to collect during testing

- Time to complete a task.
- Time to complete a task after a specified time away from the product.
- Number and type of errors per task.
- Number of errors per unit of time.
- Number of times online help and manuals accessed.
- Number of users making an error.
- Number of users successfully completing a task.

How to capture data



- Cameras
  - Point-of-View
- Software

- Eye tracking
  - Tobii
  - Gazepoint

- Remote
  - UserZoom

Field studies

- Field studies are done in natural settings.
- “In the wild” is a term for prototypes being used freely in natural settings.
- Aim to understand what users do naturally and how technology impacts them.
- Field studies are used in product design to:
  - identify opportunities for new technology;
  - determine design requirements;
  - decide how best to introduce new technology;
  - evaluate technology in use.

Field Studies - Data collection & analysis

- Observation & interviews
  - Notes, pictures, recordings
  - Video
  - Logging

- Analyzes
  - Categorized
  - Categories can be provided by theory
    - Grounded theory
    - Activity theory

Post-tests Involving Users

- Interviews
  - Get users feedback on what they have evaluated
  - Does what they are saying match how they performed?

- Questionnaires

System Usability Scale (SUS)

- The SUS is a 10 item questionnaire.
  - I think that I would like to use this system frequently.
  - I found the system unnecessarily complex.
  - I thought the system was easy to use.
  - I think that I would need the support of a technical person to be able to use this system.
  - I found the various functions in this system were well integrated.
  - I thought there was too much inconsistency in this system.
  - I would imagine that most people would learn to use this system very quickly.
  - I found the system very cumbersome to use.
  - I felt very confident using the system.
  - I needed to learn a lot of things before I could get going with this system.

Standardized User Experience Percentile Rank Questionnaire (SUPR-Q)

Usability

- This website is easy to use.
- It is easy to navigate within the website.

Credibility (Trust, Value & Comfort)

- The information on the website is credible.
- The information on the website is trustworthy.
- *I feel comfortable purchasing from this website. (Alternate for eCommerce websites)*
- *I feel confident conducting business with this website. (Alternate eCommerce websites)*

Loyalty

- How likely are you to recommend this website to a friend or colleague?
- I will likely visit this website in the future.

Appearance

- I found the website to be attractive.
- The website has a clean and simple presentation.

Interpreting Data

- **Reliability:** does the method produce the same results on separate occasions?
- **Validity:** does the method measure what it is intended to measure?
- **Ecological validity:** does the environment of the evaluation distort the results?
- **Biases:** Are there biases that distort the results?
- **Scope:** How generalizable are the results?

Evaluation and Reporting of the data

- What is the data actually saying?
- What are the common themes between participants?
- How can the system be improved?

- Typical reports provide a list of recommendations for the developers on how to improve the system.

What does this data tell you?

	Playing against computer		Playing against friend	
	Mean	St. Dev.	Mean	St. Dev.
Boring	2.3	0.949	1.7	0.949
Challenging	3.6	1.08	3.9	0.994
Easy	2.7	0.823	2.5	0.850
Engaging	3.8	0.422	4.3	0.675
Exciting	3.5	0.527	4.1	0.568
Frustrating	2.8	1.14	2.5	0.850
Fun	3.9	0.738	4.6	0.699

Table 13.1 Mean subjective ratings given on a user satisfaction questionnaire using a five-point scale, in which 1 is lowest and 5 is highest for the 10 players. Identifying strongly with an experience state is indicated by a higher mean. The standard deviation indicates the spread of the results around the mean. Low values indicate little variation in participants' responses, high values indicate more variation

Data presentation

- The aim is to show how the products are being appropriated and integrated into their surroundings.

- Typical presentation forms include:
  - Vignettes
  - Excerpts
  - Critical incidents
  - Patterns
  - Narratives

Subject Outline - Reading

- Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing user satisfaction in the era of user experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. *International Journal of Human-Computer Interaction*, *31*(8), 484-495.

Next Week

- Usability Evaluation Methods II
  - Predictive Evaluations
    - Inspection methods
  - Testing in practice

Take Home Message…

What should you test/when and with who?

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Questions

UNIVERSITY
OF WOLLONGONG
AUSTRALIA