Evaluation of Interactive Systems

Observation:
Think aloud method and User Experiments

Some sources used in this class:
Saul Greenberg's class
Wendy Mackay's class

Caroline Appert - 2018/2019
Method for testing usability
Observation

Measuring the usability of a system/technique X by observing the user while he/she uses X

Involve the user → Ethics concerns

More or less controlled

Real environment / lab

Short-term / Long-term
Ethics

Testing can be a distressing experience

- pressure to perform, errors inevitable
- feelings of inadequacy
- competition with other participants

Golden rules

- participants should always be treated with respect
- always explain you are testing the system, not the user
Ethics - Before the test

Don’t waste participant’s time
  debug and set up the experiment environment

Make participants feel comfortable
  acknowledge that the software may have problems
  let participants know they can stop at any time

Maintain privacy
  tell participants that individual test results are confidential
  explain any monitoring that is being used

Only use volunteers
  participant must sign an informed consent form
Ethics - During the test

Don’t waste participants’ time
   never have participants perform unnecessary tasks

Make participants feel comfortable
   try to give participants an early success experience
   keep a relaxed atmosphere in the room
   coffee, breaks, etc
   deliver test tasks one at a time
   never indicate displeasure with participants’ performance
   avoid disruptions
   stop the test if it becomes too unpleasant (gracefully)
Ethics - After the test

Make participants feel comfortable
state that the participant has helped you find
areas of improvement

Inform the participant
answer particular questions about the experiment
that could have biased the results before

Maintain privacy
never report results in a way that individual
participants can be identified
only show videotapes outside the research group
with participants’ permission
Method for testing usability
Observation

Think aloud protocol
Activity loggers
Lab experiments
Quasi experiments
Field study (not developed in this class)
Think aloud protocol
What it is

Ask users to express out loud what they are looking at, thinking and doing while using the interactive system and record (note taking, audio/video recording)

Can be

Task-oriented: the evaluator asks the user to perform a specific task with the system

Open-ended: no task is specified, and the user is free to choose their own task
Think aloud protocol
Relation evaluator / participant

The evaluator does not help the participant...

...but the participant can get stuck. In that case, the evaluator:

tries first to prompt without biasing (e.g., “What are you thinking about?”, “Why did you do that?”)

if it does not help, the evaluator helps the participant but carefully notes what he/she said and what happened afterwards to analyze the effect of his/her input.
Think aloud protocol
Pros and cons

Pros

Rapid, discount, only one evaluator

Cons

Not natural, oral communication is not as fast as thoughts and acts (unless a specific problem arises)

Ccl: Think aloud is an evaluation technique of limited use
Method for testing usability
Observation

Think aloud protocol

**Activity loggers**

Lab experiments

Quasi experiments

Field study (not developed in this class)
Activity logger

Log users’ actual activity using a program

Only access to low-level input (input devices)

- closed applications developed by third-party
  ==> not very useful, difficult to interpret

Access to the interface’s semantics (commands…)

- applications developed internally, pluggable API, open source, …
  ==> can be used to picture users’ contexts and activities
From low-level to high-level

Some techniques can be used to translate low-level actions (mouse clicks, key presses, etc.) into high-level actions (Ok button, undo, etc.)

Using image-analysis techniques

Using accessibility API

But they are still experimental

[Dixon & Fogarty, CHI ’10]
Detecting usability problems

Without knowing users’ tasks at hand, even high-level commands do not tell that much

example: the user has left a website. Was he only browsing or did he fail at buying something?

Getting more than broad statistics on usage (e.g., command X has been used 50 times for a 10-hour use) usually require to ask users to follow some instructions that set a context for interpreting logged activity
Detecting usability problems

Activity logging cannot reveal all kinds of usability issues

e.g., problems related to feature discoverability

Undo, Cancel and Erase events can tell a lot

example: study with 35 participants with Google SketchUp [Akers et al., CHI ’09]

video-record participants’ activity

classify exploratory activity vs. critical incidents

show them clips around undo and erase events and ask them to comment on them
Method for testing usability
Observation

Think aloud protocol
Activity loggers

Lab experiments
Quasi experiments
Field study (not developed in this class)
Experiment - definition

“A test under controlled conditions that is made to demonstrate a known truth, examine the validity of a hypothesis, or determine the efficacy of something previously untried”
Lab experiment

Propose and verify a research hypothesis in a controlled lab setting

Users point faster with a mouse than with a trackpad

Gesture commands are easier to recall than keyboard shortcuts

Users make more typing errors with software keyboards than with physical keyboards

...
Hypothesis

A supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation

A hypothesis should be:

*testable*: the means for manipulating the variables and/or measuring the outcome variable must exist

*falsifiable*: must be able to disprove the hypothesis with data

*precise*: should be specific (operationalized)
Testing a hypothesis

The experimenter manipulates **independent variable(s)** (factors) and collects **dependent variable(s)** (measures)

---

**Independent variable**
what is manipulated

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>relationship</th>
<th>Variable 2</th>
</tr>
</thead>
</table>

**Dependent variable**
what is measured
Manipulable cause

An independent variable is a potential cause that can be *manipulable*

- the dose of a medicine is manipulable
- genetic material is not manipulable

True experiments can explore the effects of things that can be manipulated

That does not mean that a non-manipulable thing cannot be a cause! That just means that experimentation does have limits.
Falsifying a hypothesis

Statistics tests on observed data can falsify a hypothesis, it can not validate a hypothesis.

If measured data is coherent with a hypothesis, the experimenter can not conclude anything.

The experimenter research question is usually turned into the null hypothesis so as to reject it and support his research hypothesis.
Testing a hypothesis

research hypothesis: Users point faster with a mouse than with a trackpad

null hypothesis: Users point as fast with a mouse as they point with a trackpad

Independent variable

Pointing device

{Mouse, Trackpad}

leads to equal

Dependent variable

Pointing time
Testing a hypothesis

research hypothesis: Gesture commands are easier to recall than keyboard shortcuts

null hypothesis: Gesture commands are as easy to recall as keyboard shortcuts are

**Independent variable**

Shortcut modality

{Gesture, Keyboard}

**Dependent variable**

Recall

{Yes, No}

leads to equal
Testing a hypothesis

research hypothesis: Users make more typing errors with software keyboards than with physical keyboards

null hypothesis: Users are as accurate with software keyboards as they are with physical keyboards

**Independent variable**

**Dependent variable**

leads to equal

Keyboard type

{Software, Hardware}

Number of typing errors
Operationalization

Defining a precise way (i.e. a task for users to do) of observing dependent variables as a function of independent variables in an experimental task.

The task must be simplified to its minimum so as to eliminate bias and effects from confounding variables (i.e. variables that are not of interest)
Confounding variable

Any variable other than the independent variable that can possibly explain the change in measures

Learning

e.g. all participants are first tested with the physical keyboard and then with the software keyboard → software keyboard has the advantage that participants have learned the keyboard layout

Prior experience

e.g. use conventional keyboard shortcuts (e.g. ctrl+V for paste) when comparing them to gesture shortcuts, which are a non-familiar type of shortcuts → keyboard shortcuts are favored because of participants’ existing knowledge
Operationalization

Operationalization consists of turning a phenomenon into an experimental task that takes factors as input and outputs measures.

“Children grow more quickly if they eat vegetables”

Meaning of ‘children’? (e.g. 4 < age < 8)
Meaning of ‘grow more quickly’? (e.g. cm per year)
What are ‘vegetables’? (e.g. quantity of vitamin C)
Validity issues

Validity

best available approximation to truth or falsity of propositions

Internal validity

best approximate truth about inferences regarding causal relationships (i.e. effects observed are attributable to controlled variables)

External validity

generalizability of the results to other subjects and situations

Operationalizing requires to find the good trade-off internal validity - external validity
Causality and Correlation

Correlation

mathematical relationship between two variables.

Causality

physical relationship between two variables. There is a chain of events when the first variable varies that causes the other variable to vary (involves time).
Causality and Correlation

Correlation does not imply causality

For example, we noted a high correlation between height and weight. This does not show that height \( \Rightarrow \) weight.

"When does correlation imply causation?"

When the experiment design is done with appropriate care to avoid confounding and other threats to the internal validity of the experiment.
Operationalization

Well-known standards

If some well-known standards do exist, use them

Pointing

ISO. 9241-9 Ergonomic requirements for office work with visual display terminals (VDTs)-Part 9: Requirements for non-keyboard input devices.

Text entry

MacKenzie et al.’s phrase set

http://www.yorku.ca/mack/chi03b.html
Operationalization
Well-known standards - Pointing

The circular layout and order of appearance of targets force participant to point in every direction
Operationalization
Well-known standards - Text entry

Phrases that are moderate in length, easy to remember, and representative of the target language

- video camera with a zoom lens
- have a good weekend
- what a monkey sees a monkey will do
- that is very unfortunate
- the back yard of our house
- I can see the rings on Saturn
- this is a very good idea
- ...
Operationalization
No standard - Example#1

Learning Keyboard vs Gesture shortcuts

use arbitrary mappings (avoid the effect of prior experience)

use a number of shortcuts representative of expert usage (e.g. 14)
Operationalization
No standard - Example #1

Learning Keyboard vs Gesture shortcuts

measure recall on two consecutive days
(learnability + memorability)
Operationalization
No standard - Example#2

MultiScale searching in zoomable interfaces using different navigation techniques

Ask participants to inspect $n$ objects until they found the target (i.e. the square that has rounded corners).
Operationalization
No standard - Example#2

MultiScale searching in zoomable interfaces

control chance bias:

force a quantity of exploration a priori
(at least n objects to inspect) with software control

control individual differences in visual acuity:

add an explicit action to unveil corners
(press space bar) that is available only at a scale
where all participants can perceive rounded corners
(need pilot studies)
Dependent variable (measure)
Types of scales

Four major scales of measurements
- Nominal
- Ordinal
- Interval
- Ratio

Possible manipulations depend on the type of scale
Dependent variable (measure)
Nominal scale

Classification into named or numbered unordered categories
  country of birth, user groups, gender…

Possible manipulations
  whether an item belongs in a category
  counting items in a category
Dependent variable (measure)
Ordinal scale

Classification into named or numbered ordered categories (no information on magnitude of differences between categories)

- preference, social status, gold/silver/bronze medals

Possible manipulations

- as with nominal scale, plus
- merge adjacent classes
- transitive: if A > B > C, then A > C
Dependent variable (measure)
Interval scale

Classification into ordered categories with equal differences between categories (zero only by convention)

- temperature (C or F), time of day

Possible manipulations
- add, subtract
- cannot multiply as this needs an absolute zero
Dependent variable (measure)
Ratio scale

Interval scale with absolute, non-arbitrary zero length, weight, time periods

Possible manipulations
multiply, divide
Types of design

Choose a representative sample of the population you want to study.

How to assign the different levels (i.e. values) of an independent variable?

Which participants will test pointing with a mouse and/or pointing with a trackpad?

Which participants will test keyboard and/or gesture shortcuts?
Types of design
Between-subject design

If the independent variable has n levels

Randomly select n groups and assign a different level to each group

Assumptions

Other non-controlled variables are randomly distributed between the n groups

The only systematic difference between the n groups is the independent variable
Types of design
Within-subject design

If the independent variable has n levels
Successively expose each participant to the n different levels
Automatically control of most of the other variables
Allows experimenter to use a smaller number of subjects
Types of design
Factorial design

Test several independent variables in the same experiment.

Mouse vs. Trackpad: device (mouse, trackpad), target width (10, 20, 30), target distance (200, 400)

Each variable can be distributed according to a between or within subject design.

A combination of a value for each independent variables defines an experimental condition

2 device x 3 target width x 2 target distance = 12 conditions
Controlling variation

Replication and blocking are two mechanisms to reduce observed variation that is not due to difference between conditions

Replication: A participant does several times the experimental task in the same condition.

  e.g. if the participant got distracted in a particular condition

Blocking: Arranging the tasks into blocks of tasks that are similar to one another.

  e.g. eliminating the time due to successive changes between two pointing devices
Controlling order effect

Order effect happen when an independent variable is presented according to a within-subject design

if the mouse is always presented after the trackpad and observed time is shorter with the mouse, is pointing performance better because of the input device or because the user has become more familiar (thus efficient) with the task?

Randomization: presentation the different conditions in a “random” order across the experiment
Randomization

Randomization is not haphazard

An experiment is randomized if the method for assigning levels of independent variables involves a deterministic probabilistic scheme.

Examples of bad randomization

assign mouse to the first half of participants who show up and trackpad to the other half (pb: the first half can be the most motivated participants)

assign mouse or trackpad depending on if start time in seconds is a odd or even (pb: can result in much more observations in one or the other condition)
Counterbalancing

Counterbalancing is a scheme to randomize a within-subject experiment design.

Consider an independent variable that has $n$ levels and a sample of $X$ participants.

- Compute the $n!$ possible orders and assign each order to $X / n!$ participants (requires a potentially high number of participants).
- Compute $n$ possible orders using a Latin Square and assign each to $X / n$ participants.
Randomization
Latin square definition

A Latin square is an $n \times n$ array filled with $n$ different symbols, each occurring exactly once in each row and exactly once in each column.

Example: $n=3$ levels ({A, B, C})
Method for testing usability

Observation

Think aloud protocol
Activity loggers
Lab experiments

Quasi experiments
Field study (not developed in this class)
Quasi-experiments

True Experiment
  random assignment of units to conditions

Quasi Experiment
  units are not assigned to conditions randomly

Quasi-experiments are very similar to true experiments but use naturally formed groups (sometimes called natural experiments)
Quasi-experiments

When randomization is not possible or unethical

- e.g. Spanked children are less responsive to an educational class for reducing aggressiveness (ethics)
- e.g. Vaccine X is ineffective for people having the gene Y (not possible)

External validity may be higher as quasi experiments can be seen as “more natural”

Internal validity may be lower as other differences between groups may exist
Quasi-experiments

Make explicit the “threats of internal validity”
  history, selection of participants, participant loss, etc

Increase internal validity using design strategies
  pretest-posttest
  Interrupted time-series...
Designs for Quasi Experiments

Pretest-posttest

Measure before and after manipulating independent variables

Interrupted time-series

Several measures before and after manipulating independent variables
Captures trends → higher internal validity
Replicability

Any experiment should be replicable by others

Always report:

- a complete description of the experiment design
- the hardware/software characteristics of the experimental environment (apparatus)
- a description of the participants’ characteristics that may impact the observed measures (gender, mean and variance in age, prior experience, social and occupational aspects...)

Running the experiment

Control for bias, avoid Hawthorne effect

Hawthorne effect: changes in participants' behavior during the course of a study may be "related only to the special social situation and social treatment they received."

unbiased instructions (write them down before)

double-blinded if possible (the experimenter and the participant do not know which group it is)

Keep the same conditions (software, environment) from one participant to another
Laboratory vs Crowdsourcing

Laboratory

small pool of participants (10 to 40 participants)

supervised setting

Crowdsourcing

large pool of participants (up to thousands participants)

unsupervised setting
Crowdsourcing

What for

Participants complete a micro-task remotely and are compensated with micro-payments (e.g., Amazon Mechanical Turk)

Used for tasks that require human intelligence such as identifying objects in images, finding relevant information, doing natural language processing

Suitable to HCI experiments, especially for performance-based evaluations (“do as fast as possible” instruction)
Crowdsourcing Threats

Crowdsourcers can

give random responses (malicious behaviors),
be interrupted (e.g., phone call, watch TV, etc.),
have varying devices and browsers,
have disabilities,
...

Crowdsourcing Threats
Crowdsourcing
Good practices

Ensure instruction comprehension using qualification tasks before running the experiment

e.g., use a multiple choice question where only one answer is correct

Consider age and hardware to allocate participants to conditions, especially for between-subject designs
Crowdsourcing
Good practices

Encourage honest reporting of problems

e.g., ask for reporting any disability or technical difficulty that might have impacted performance while assuring that participants will still receive compensation

Clean collected data using a severe outlier detection strategy

e.g., remove any measure that is more than 3 x IQR higher than Q3, or that is more than 3 x IQR lower than Q1

Interquartile Range
IQR = Q3 - Q1
Recording dependent variables

Log data

Save a file that is easy to analyze by man and machine

detail each acronym/short name you may use for denoting a variable or a level (e.g. TP means TrackPad)

log the run date

tabular format with one line per run task (trial)

Collect independent variables, dependent variables and any data that could be useful to “replay” the experiment.
Recording dependent variables

Log data

Collect computed measures and raw data

to fix potential undetected bugs

e.g. in an experiment testing the accuracy of a gesture recognizer, collect all (x,y,t) and recognized gesture class. This allows you to a posteriori fix the recognizer if needed without re-conducting the experiment.

to allow you to do analyses that you did not plan in advance

e.g. in a pointing experiment, collect (mouse_x,mouse_y, target_x, target_y) and hit/miss. This allows you to get an estimation of the distance error even if you did not plan it.
Experiment design exercise 1
Multitask switching techniques

Design an experiment to evaluate the Pile and List techniques

Hypothesis(es)
Factors
Measures
Task

Pile
List
Experiment design
exercise 2
Caret positioning techniques

Design an experiment to evaluate the semi-direct vs indirect techniques

| Hypothesis(es) | semi-direct: dwell on text to make a slightly offset magnifying glass appear |
| Factors        | indirect: dwell on keyboard to turn it into a trackpad |
| Measures       |                                                       |
| Task           |                                                       |