Software Development Methodologies

Lecture 9 - User Studies 2

SOFTENG 750 2013-05-01

Murphy's Law for Experimentalists

Anything that can go wrong will go wrong.

1. If something can go wrong, it will do so just before your deadline.
2. If the reading on your detector is correct, then you forgot to plug it in.
3. If several things can go wrong, then they will do so all at the same time.
4. If nothing can go wrong with your experiment, something still will.
5. If you make a great discovery today, you will find a major error in your methods tomorrow ("here today, gone tomorrow").

...

Edward A. Murphy, Jr.

Learning Outcomes

You will be able to...

1. Create your own study designs
2. Write scripts for facilitators

3. Anticipate and avoid sources of errors

4. Describe empirical results

SE Software Engineering The University of Auckland

User Study Design

Recap: Example Design of a Usability Study

Based on your research question, you define...

1. **Independent Variable**
   User Interface (*UI*)**,** two different levels: *old* and *new*
2. **Dependent Variables**
   performance, *accuracy*, *satisfaction*
3. **Hypotheses**: what outcome do we expect
   "*new* has better performance, accuracy & satisfaction"
4. **Tasks**: what do the participants do? "accomplish goal X"
5. **Measurement**: how do we measure the dependent variables?
   Task completion time, count mistakes, questionnaire at the end (*demographics* & *satisfaction*)
6. **Schedule**: who does what, when, how often?
   Group1: 3 tasks with *old*, then 3 tasks with *new*
   Group2: 3 tasks with *new*, then 3 tasks with *old*

Threats to Validity I

Misunderstandings by participants or facilitators:

- What to do? How? What to record? How to measure?
- Define it exactly in the *script*

Order Bias: the order of the tasks has an effect on the DVs

- Permute task order to distribute the order bias equally
- You can analyze later if there is

Training Effect: participants get better the more tasks they do

- Add training phase: training tasks before each type of task
- Permute task order to distribute training effect

Fatigue: participants get tired after hard tasks, performance loss

- Schedule breaks (between the tasks / after *n* tasks)
- Permute task order to distribute effect of fatigue

Threats to Validity II

Social Desirability Bias (als Acquiescence Bias):

- Participants tend to do what is socially expected (they are nice, show respect & support, don't criticise openly)
- Participants may be your friends who want to support you
- Make clear honest results are most valuable
- Make it less personal, e.g. written instructions, facilitator not looking at questionnaire answers

(Self-)Selection Bias

- The people (who choose to) participate in your study may be special (e.g. only tech enthusiasts) and not representative
- Try to diversify your group of participants, e.g. advertise differently

Confounding Variable:

- Variable that is not controlled, but has an effect
- E.g. comparing UIs A and B, but testing A on a larger screen than B (screen size is the confounding variable here)
- Is the result due to the IVs or just the confounding variable? Control it, i.e. use same screen for all conditions

Creating a Script

- Write down **step-by-step** what the facilitator should do
  - Some steps are vital for measurement
    (e.g. when to take time, what to note down)
  - Inconsistencies between sessions can disturb the results!
- Consider **possible cases and exceptions**:
  - **When to help?** have a clear procedure in place
    - It may be ok to provide help, e.g. answer questions
    - But do this consistently & make sure participants know
  - **When to stop?** (if participants can't complete the tasks)
  - What if participant **wants to stop**? they can anytime
- Estimate **how long** each session takes (may need to adjust)
- How will the **data** be collected?
  Have procedure in place, e.g. facilitator's log

Example Script Part I
Comparing Systems A and B

1. **Preparation**: start software A and B, maximize windows
2. **Greeting**: welcome & briefly introduce project
3. **Ethics**: let the participant read the **information sheet** and sign the **consent form** (collect the form)

General procedure for each condition:

4. **Training**: Explain & walk through task 1, ensure that participant is confident with it and all questions are answered
5. **Execution**:
   1. Make sure system is in start condition and ready
   2. Give participant instructions for task 2
   3. Confirm that task is understood & ready, start taking time
   4. Note down significant observations during task
   5. Stop & note down the time once task goal is reached
   6. Small break, then repeat with task 3

Example Script Part II
Comparing Systems A and B

Scheduling the tasks:

- Prevent order bias by varying the order (AB or BA)
- Odd-numbered participants start with system A:
  - First A: training task 1, then tasks 2 and 3
  - Then B: training task 4, then tasks 5 and 6
- Even-numbered participants start with system B:
  - First B: training task 1, then tasks 2 and 3
  - Then A: training task 4, then tasks 5 and 6

Wrapping up:

1. Post-task questionnaire:
   Demographics & satisfaction questions for A and B
2. Dismissal

Conducting the Study

Pilot Study with about 3 participants

- Good practice for facilitators
- Refine the design & script
- Get an initial feeling for the results to come

Main Study, typically with 30+ participants

- No drastic changes unless really necessary
- Several facilitators may work in parallel if possible
- Ideally all data collected in one spreadsheet (e.g. transcribe data from logs & questionnaires)

Data Analysis and Publication (if results valid & interesting)

Collecting and Analyzing Data

Scale Type of a Variable

Each variable has a scale type that tells us what we can do with the variable values, i.e. which operations make sense.

Scale Type	Examples	Important Operations
Nominal (only categories)	Different systems (e.g. UIs) Participant comments Phone numbers	=, ≠, mode, frequency
Ordinal (values can be ordered)	Rankings (e.g. preference) Some questionnaire scales School grades	All from nominal plus <, >, median, percentile
Interval (differences are meaningful)	Some questionnaire scales (e.g. standard Likert-scale), Temperature in C or F	All from ordinal plus -, average, standard deviation
Ratio (ratios are meaningful)	Time Distance Counts (frequencies)	All from interval plus /

Data Spreadsheet

- Ideally all data is collected in a single spreadsheet
- One **row** for each participant
- One **column** for...
  - Participant number
  - Scheduling (e.g. A first or B first)
  - Each DV for each task for each condition, e.g. time/errors for task 1..6 in conditions A & B = 24 columns
  - Each questionnaire question (e.g. demographics, Likert-scale items, -2 to +2)
  - Observations & Comments

Num	First	t1A	t1B	t2A	t2B	t3A	t3B	t4A	t4B	Q1	Q2	Age	Comments
1	A	12	---	9	---	---	80	---	30	-1	2	25	Crash in 3
2	B	---	98	---	30	19	---	46	---	-2	2	21	Didn't get 1
3	A	5	---	13	---	---	13	---	23	0	1	23	Cheated in 1
4	B	---	10	---	27	20	---	93	---	-1	2	30	Fell asleep

Describing Results 1

Interval & Ratio Variables

1. Describe your data by **aggregating** it in the spreadsheet
2. Calculate **average** and **standard deviation** for columns
    - ○ Avgs show the general trend
    - ○ Std devs show how much values are spread out (-> how much do they differ with each other, precision)
3. Calculate the **average differences** between the DVs in different conditions, e.g. avg(t1A) - avg(t1B)
4. Create **summary table** showing only the avgs, std devs and average differences

Variable	**Average (seconds)**	**Std. Dev.**	**Difference between averages for A and B**
t1A (time for task 1 on system A)	20	2.5	
t1B (time for task 1 on system B)	15	2.8	5
t2A (time for task 2 on system A)	36	5.8	
t2B (time for task 2 on system B)	28	6.2	8

Describing Results 2

Nominal Variables (from open questions, observations, etc.)

1. Categorize the answers, count how often each answer was given (**counts / frequencies**)
2. Sort answers descendingly by frequency
3. Create **summary table** showing answers and frequencies

What did you dislike about B?	**Frequency (out of 15)**	**%**
System B was too slow.	5	50
System B didn't have function X.	3	20%
System B is not as flash as A.	2	13%
System B is hard to understand.	2	13%

Causality between Vars

"A changes together with B" (correlation) could mean...

1. A causes B
2. B causes A
3. A and B are caused by another variable C
4. Any comibination of the above

Controlled Studies:

- Changes of IVs assumed as cause for changes of DVs
- Because we try to keep everything else the same
- Otherwise confounding variable

Observational Studies:

- Causality much harder to determine
  (because no IVs, possibly many confounding variables)
- Example: ice cream and drowning

Today’s Summary

1. Empirical studies need the following preparations:
   - ○ **Design**: IVs and DVs, tasks, measurement
   - ○ **Script**: step-by-step guide for facilitators
2. **Pitfalls** such as possible misunderstandings and biases need to be considered
3. **Pilot study** to iron out the problems early on

Milestone (Deadline: Lab on Thursday)

1. Design a usability study for your app
2. Create a script for the study

Quiz

1. What is a confounding variable?
2. What is training effect and how to mitigate it?
3. Why do we perform pilot studies?

```
word_to_anagram = "documenting"

def remove_trailing_whitespace(text):
  return text.rstrip()

word_list = [ remove_trailing_whitespace(line) for line in file("wordlist.txt")
]

for first_word in word_list:
  for second_word in word_list:
      if sorted(word_to_anagram) == sorted(first_word + second_word):
          print word_to_anagram, "=", first_word, "+", second_word
```

Self-Documenting Code Contest 2008
Winning entry by Ian Davis
http://selfexplanatorycode.blogspot.co.nz/