COMPSCI 715

Empirical Studies

If we knew what it was we were doing, it would not be called research, would it?
(Albert Einstein)

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.

Empirical Studies

You have solved a problem
or answered a research question.

How do we know your solution is useful
or your answer true?

Solution: conduct an empirical study!

1. Choose a methodology to investigate research questions
2. Use methodology to collect data
3. Interpret data to find answers and verify if hypotheses

Often the only way to convince people!

Qualitative vs Quantitative Studies

Qualitative Study

- Show users a (paper-) prototype and get feedback
- Possibly let them use the prototype (e.g. specific tasks)
- Data are typically text:
  interviews, open questions, think-aloud protocol, observations, ...
- Good to explore an early prototype:
  What are the problems with this UI?

Quantitative Study

- Data are numbers: tasks or questionnaires with measurable outcomes
- Rigorous methodology:
  variables, hypotheses, measurements, statistics...
- Good to compare solutions:
  Which one is better? How much better?

Often mixed-methods approach: quantitative & qualitative

How to Conduct an Empirical Study

1. Choose research questions
   - Specific enough to be answerable, general enough to be interesting
   - Specify the target population

2. Specify methodology
   - Define tasks (i.e. what do users do during the study)
   - Define independent and dependent variables and specify how they are measured
   - Specify hypotheses based on the variables
   - Create a script for the study (step-by-step guide)

3. Conduct a pilot study and revise methodology

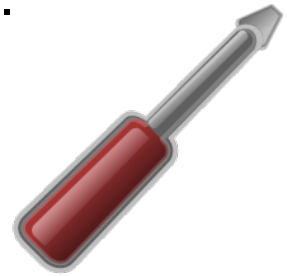
4. Use script to collect data (e.g. from participants)

5. Analyze the data, test hypotheses, interpret & discuss

Variables

Independent Variable (IV): What do I change?

- Variable of which we want to know the effect:
  we change it (try different values) and see what happens
- Levels: the different values that we try out & compare
  e.g. the UI used, VR vs screen, different mechanics, ...
- Levels lead to conditions that need to be tested
  - Usually only few independent variables and levels
    - More conditions means more time required



Dependent Variable (DV): What do I observe?

- Variable that describes the effect we are investigating
- Needs to be measurable (as accurately as possible)
  - Can be many if measuring them is cheap
    (“if in doubt, measure everything”)
  - E.g. completion time, questionnaire score

Example: 3D Displays and Spatial Memory

Independent Variable: display condition

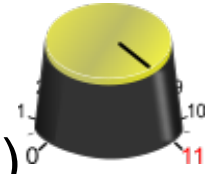
- 2D (control condition)
- 3D: all combinations of HCP, stereo-display, landmarks

Display Cond.	Stereoscopy	Landmark	Head Tracking	3D Models and Arrangement
2D	--	--	--	--
3D	--	--	--	✓
3DH	--	--	✓	✓
3DL	--	✓	--	✓
3DHL	--	✓	✓	✓
3DS	✓	--	--	✓
3DSH	✓	--	✓	✓
3DSL	✓	✓	--	✓
3DSHL	✓	✓	✓	✓

Dependent Variables: task duration, number of mistakes

Types of Empirical Studies

Controlled Study:



- Change the independent variables (try different values)
- Measure the dependent variables (to find out about effects)
- Keep everything else the same as much as possible

Observational Study:

- Variables are not controlled, but merely observed
- Try to infer effects from the observed values
- Sometimes necessary because variables can be difficult to control (e.g. weather, user behavior in big organization)

Lab vs Field Study:

vs

- Lab: More control (good for controlled studies), less "contamination" by uncontrolled variables
- Field: Less control (i.e. usually observational), more realism

Requirements

Ethics Approval: is the study ethical?

- Most big organizations require approval process
- Regulate risks: damage, power abuse, deception...
- Standard practice:
  give participants info sheet to read,
  and consent form to sign

Participants: how to recruit people from the target population?

- Advertise in the right places (often low response rate)
- Motivation: share results, reward (money, voucher, food etc.)
- Try for a representative sample, e.g. gender balance

User Study Design

Defining Tasks

Tasks of a controlled user study should be...

- **Relevant** for answering the research question
- **Well-defined**: it should be clear for a participant what to do
  - Have a clear starting point, clear goal
  - Avoid choices, otherwise the data gets erratic
  - Make it easy to train unfamiliar participants
- **Realistic**: how do real users do it?
  - Make results generalizable to the real world
  - Find balance with well-definedness/simplicity
- **Variations**: find similar tasks that test exactly the same thing
  - Need tasks for training, different levels of IV, repetitions to generalize results to a whole use case (not just the task)
- **Well-timed**: how long does the task take?
  - Prevent data gathering from getting too time consuming
  - Ensure task doesn't get out of hand (consider worst case)

For qualitative & observational studies task definitions less important

Measuring Usability

Performance

- *Task completion time*, operation counts, *eye gaze path length*
- Specific performance scores*:* e.g. productivity scores

Accuracy

- Number of mistakes: define what exactly counts as a mistake
- Or measure deviation from an objective optimum
  (e.g. "align objects perfectly", "find cheapest price")

Satisfaction

- Subjective but often more important than performance
- Typically measured with questionnaire
  (e.g. Likert-scale, "I enjoyed using the system")

Measuring DVs

Define precisely how you measure your dependent variables.

Quantitative Measures

- *Task completion time*: define start and finish events
- *Event counts:* key strokes, mouse clicks, mouse path, keyboard/mouse switches (usually from event log), mistakes
- *Eye gaze path length* (from eye tracker)
- 5-point Likert-scale items with standard labels (subjective!) "strongly disagree" to "strongly agree"
  - Common Likert-scale labels: http://www.gifted.uconn.edu/siegle/research/Instrument%20Reliability%20and%20Validity/Likert.html
  - http://dataguru.org/ref/survey/responseoptions.asp

Qualitative Measures

- Open questions, e.g. "What did you dislike about the system?"
- Think-aloud protocol statements
- Observations, e.g. observed participant comments/reactions

System Usability Scale

Measures subjective usability with standard 5-point Likert scale:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

See also http://www.measuringusability.com/sus.php

Measuring Demographics

Make sure to collect **demographic data**

- Data describing your sample of users (e.g. gender, age, occupation)
- They may have an effect on the DV
- Effects on the DV can be analyzed later (very interesting, e.g. gender differences)

Important to discuss whether the results can be generalized to the whole population of interest

Use questionnaire:

1. Age, Gender, Occupation
2. Relevant experience (specific to your tasks), e.g.
   - Have you used a similar application before? How much?
   - How many hours weekly do you spend playing computer games?
   - "I do a lot of 3D modeling in my everyday computer use" (Likert-

Example Study 1 - Qualitative

1. **Research Questions:**
   "What are usability problems of our app?"
   "Which features are the most important?"

2. **Tasks**: what do the participants do? "accomplish goal X"
   - At least one task for each of the features

3. **Measurement**
   - Observe mistakes made during tasks
   - Record think-aloud protocol
   - Questionnaire at the end
     (*demographics, satisfaction, ranking of features*)

4. **Schedule**: who does what, when, how often?
   - Possibly let participants decide what tasks they do
   - Possibly change task order between participants

Example Study 2 - Quantitative

1. **Independent Variable**
   Two apps: your app (*A*) and competitor (*B*)
2. **Dependent Variables**
   performance, *accuracy*, *satisfaction*

3. **Hypotheses**: what outcome do we expect
   "*A* 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 *A*, then 3 tasks with *B*
   Group2: 3 tasks with *B*, then 3 tasks with *A*

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 combination 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

Using Statistics

Descriptive Statistics:
Help to illustrate and explore data (facilitate understanding), e.g.

- Central tendency: average, median, …
- Spread of data: range, variance, …

Inference Statistics:

- Testing hypotheses with statistical tests
- How likely is it that a difference in the data is just coincidence?
- Different tests for different conditions
  - Scale type of variables (e.g. ordinal or interval)
  - Sample size (can use more powerful tests with larger size)
  - Distribution of data (e.g. normally distributed data easier to test)
  - Variances of several samples (are they the same?)

Two popular tests:

- t-Test
- Wilcoxon-Signed-Rank test