User Centered Design

[08] Late stage user evaluations

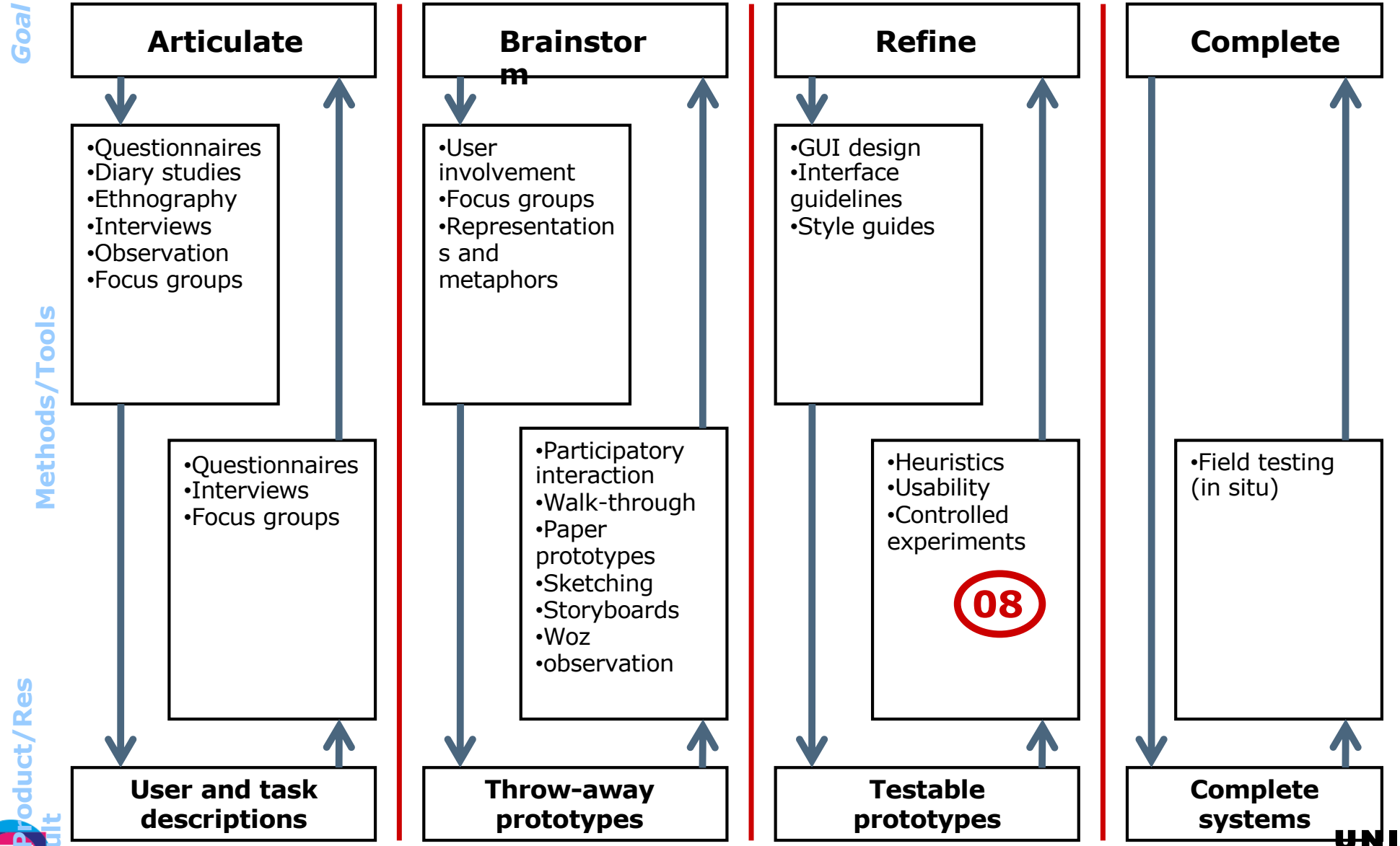
Denis Lalanne
Human-IST Institute, University of Fribourg
Nov. 20th, 2018

HUMAN-IST
HUMAN CENTERED INTERACTION SCIENCE & TECHNOLOGY

Slides inspired from S. Greenberg, K. Andrews, J. Lazar et al., and my own stuff

UNI FR

UCD Design Process & Lectures



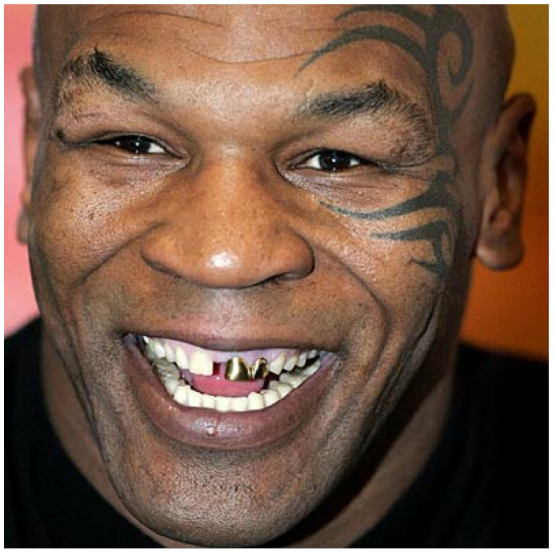
Evaluating interfaces with users

- Basic idea: directly involve people in the evaluation
  - They know their domain (usually better than you!)
- Type
  - qualitative
    - ✓ HOW: observe users, gather explanations and opinions
    - ✓ OUTPUT: list of findings
    - ✓ (+) ready explanation, easy solution,
    - ✓ (-) not measurable, hard to compare and track, chaotic process
  - **quantitative**
    - ✓ HOW: measure efficiency (time), accuracy (errors), satisfaction
    - ✓ OUTPUT: measures
    - ✓ (+) measurable, can be tracked, repeatable, allows comparison
    - ✓ (-) hard methods, difficult to translate in solutions (more about findings)

Controlled experiments

- Traditional scientific method (hypothesis testing, inferential statistics)
- Based on hypothesis and expressed in form of comparison between designed cases
  - Traditional example:

    “There is no difference in the number of cavities in children and teenagers using toothpaste or not when brushing daily over a one month period”



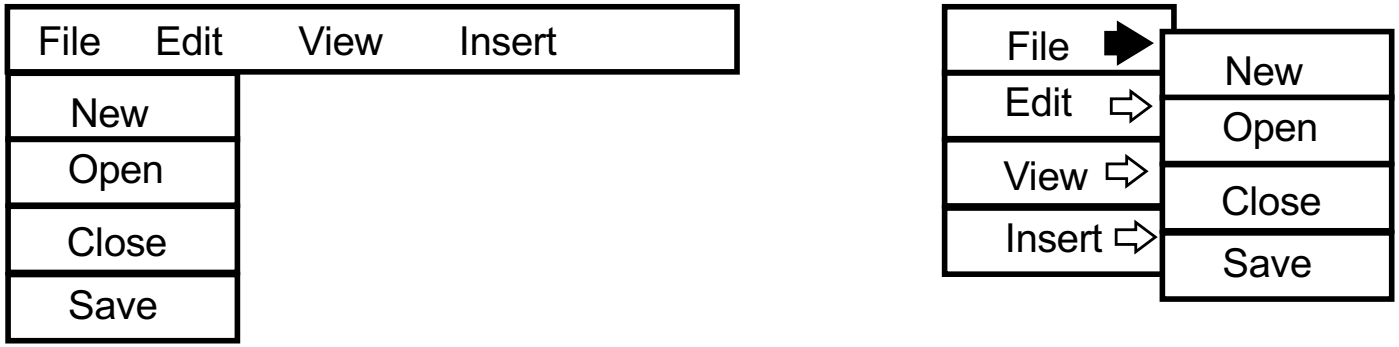
Controlled experiments

- Phases:
  - A) State a lucid, testable hypothesis
  - Define:
    - ✓ B) Independent variables
    - ✓ C) Dependant variables
  - D) Subject Selection
  - E) Controlling bias
  - F) Statistical analysis
  - G) Interpret your results

A) Clear (Lucid) and testable hypothesis

- HCI Example:

  There is no difference in user performance (time and error rate) when selecting a single item from a **pop-up** or a **pull down** menu of 4 items, regardless of the subject's previous expertise in using a mouse or using the different menu types"



A) Clear (Lucid) and testable hypothesis

- NOTE: translating high level questions to testable hypothesis is not trivial
  - examples
    - ✓ "Graphical UIs are *better* than command based UIs"
      - better in terms of what (faster, more accurate, more easily learnable)? which type of GUIs? for which kind of users? …
    - ✓ "Navigating through multiple short web pages is *better* than scrolling over one page with the same content"
      - what type of content? how many lines to scroll? how many pages to navigate? ...
  - It has strong implications over the scope of your findings
    - ✓ Tradeoff between:
      - Framing the context to a testable hypothesis
      - Generalization of observed results

B) Independent variables (IV)

- Hypothesis includes the **independent variables** that are to be manipulated
  - the things you manipulate independent of a subject's behaviour (typically interface features or competing solutions)

- *in toothpaste experiment*
  - ✓ toothpaste type: uses toothpaste or not
  - ✓ age: <= 11 years *or* > 11 years

- *in menu experiment*
  - ✓ menu type: pop-up or pull-down
  - ✓ menu length: 3, 6, 9, 12, 15
  - ✓ subject type (expert or novice)

C) Dependant variables (DV)

- Hypothesis includes the **dependent variables** that will be measured
  - ✓ The (performance) factors by which selected cases are compared
  - ✓ Variables dependent on the subject's behavior as a reaction to the independent variable
  - ✓ The specific things you set out to quantitatively measure / observe
- *Key methodological goal*
  - ➢ Single out variation dependent **exclusively** on independent (manipulated) variables
- *in menu experiment*
  - ✓ time to select an item
  - ✓ selection errors made
  - ✓ time to learn to use it to proficiency
- *Typical measures in HCI*
  - ✓ Time to complete assigned tasks
  - ✓ Number of steps required to reach a goal (e.g., mouse clicks, navigation steps)
  - ✓ Number of errors
  - ✓ Time to learn
  - ✓ Satisfaction scores
  - ✓ …

D) Subject Selection and Assignement

- How do I assign subjects to defined cases? Subject are split in groups and
  - each group assigned to a specific case (**between-group**)
    - ✓ in menu experiment
      - Group 1: pop-up
      - Group 2: pull-down
  - all subject are assigned to all cases (**within-group**)
    - ✓ in menu experiment
      - Group 1: pop-up and pull-down
      - Group 2: pull-down and pop-up
- Problem: variation in observed measures may depend on subject variability and NOT on your controlled variables
  - subjects have been split in not homogeneous groups
  - learning effects

Type of experiment design

	Between-group design	Within-group design
Advantages	Cleaner Avoids learning effect Better control of confounding factors, such as fatigue	Smaller sample size Effective isolation of individual differences More powerful tests
Limitations	Larger sample size Large impact of individual differences Harder to get statistically significant results	Hard to control leaning effect Large impact of fatigue

From Lazar et al. Research Methods in Human-Computer Interaction

D) Subject Selection and Assignement

- It is necessary to control subject variability
  - ✓ reasonable amount of subjects
  - ✓ random assignment
  - ✓ counterbalancing to deal with learning effect
  - ✓ screen for anomalies in subject group
    - – superstars versus poor performers

Novice

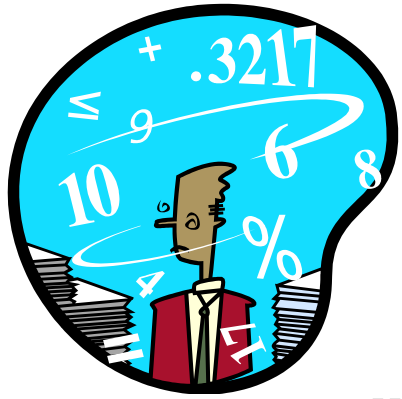
Expert

E) Controlling bias

- Control for bias
  - Take into account factors not controlled (not used as independent variables) but with potential effects on dependent variables
    - ✓ unbiased instructions
    - ✓ unbiased experimental protocols
      - prepare scripts ahead of time
    - ✓ unbiased subject selection

F) Statistical analysis

- Apply statistical methods to data analysis
  - confidence limits:
    - ✓ the confidence that your conclusion is correct
    - ✓ “the hypothesis that computer experience makes no difference is rejected at the .05 level” means:
      - a 95% chance that your statement is correct
      - a 5% chance you are wrong



G) Interpretation

- Interpret your results
  - what you believe the results really mean
  - their implications to your research
  - their implications to practitioners
  - how generalizable they are
  - limitations and critique

Statistical analysis

- Graphical analysis
  - Plot your data! Very useful as a preliminary step
  - Especially to remove outliers (more robust statistics then)
  - Scatterplots, barchats, etc.

- Calculations that tell us
  - mathematical attributes about our data sets
    - ✓ mean, amount of variance, ..
  - the probability that our claims are correct
    - ✓ “statistical significance”

Statistical vs practical significance

- When *n* is large, even a trivial difference may show up as a statistically significant result
  - eg menu choice:
    mean selection time of menu a is 3.00 seconds;
    menu b is 3.05 seconds

- Statistical significance **does not imply** that the difference is important!
  - a matter of interpretation
  - statistical significance often abused and used to misinform

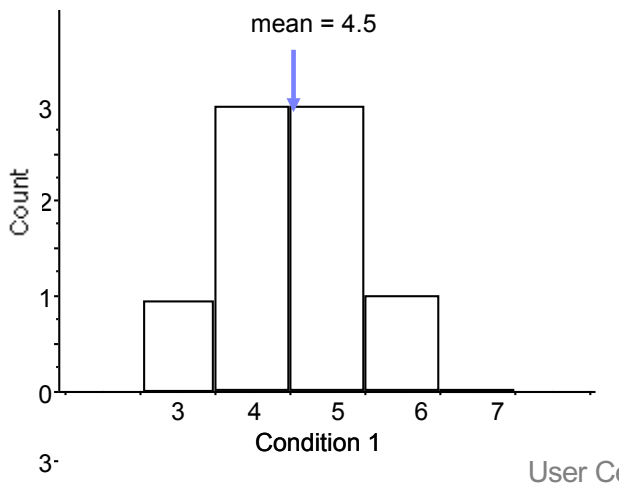
Example: Differences between means

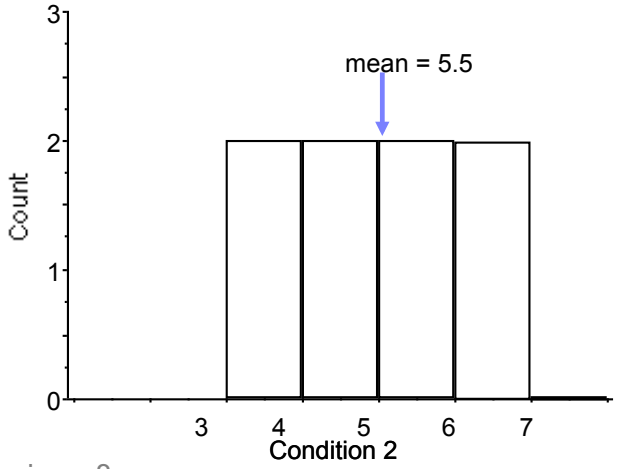
- Given:
  - two data sets measuring a condition
    - ✓ height difference of males and females
    - ✓ time to select an item from different menu styles ...

Condition one: 3, 4, 4, 4, 5, 5, 5, 6

Condition two: 4, 4, 5, 5, 6, 6, 7, 7

- Question:
  - is the difference between the means of this data statistically significant?





T-test

- A simple statistical test
  - allows one to say something about differences between means at a certain confidence level
- Null hypothesis of the T-test:
  - no difference exists between the means of two sets of collected data
- possible results:
  - I am 95% sure that null hypothesis is rejected
    - ✓ (there is probably a true difference between the means)
  - I cannot reject the null hypothesis
    - ✓ the means are likely the same

Different types of T-tests

- **Un-paired: Comparing two sets of independent observations (between-group)**
  - usually different subjects in each group
  - number per group may differ as well

Condition 1	Condition 2
S1–S20	S21–43

- **Paired observations (within-group)**
  - usually a single group studied under both experimental conditions
  - data points of one subject are treated as a pair

Condition 1	Condition 2
S1–S20	S1–S20

- **Non-directional vs directional alternatives**
  - non-directional (two-tailed)
    - ✓ no expectation that the direction of difference matters
  - directional (one-tailed)
    - ✓ Only interested if the mean of a given condition is greater than the other

Two-tailed unpaired T-test

x_1 = 3 4 4 4 5 5 5 6
x_2 = 4 4 5 5 6 6 7 7

Hypothesis: there is no significant difference between the means at the .05 level

- N: number of data points in sample
- ΣX: sum of all data points in sample
- X: mean of data points in sample
- Σ(X²): sum of squares of data points in sample
- s²: unbiased estimate of population variation
- t: t ratio
- df = degrees of freedom = N1+N2–2

$$s^2 = \frac{\sum(X_1^2) - \frac{(\sum X_1)^2}{N_1} + \sum(X_2^2) - \frac{(\sum X_2)^2}{N_2}}{N_1 + N_2 - 2}$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s^2}{N_1} + \frac{s^2}{N_2}}}$$

- In our example:
  - N1=8, N2=8
  - ΣX1=36, ΣX1=44
  - df=14
  - S²=1.1429
  - t=1-871

Level of significance for two-tailed test

x_1 = 3 4 4 4 5 5 5 6
x_2 = 4 4 5 5 6 6 7 7

Hypothesis: there is no significant difference between the means at the .05 level

- In our example:
  - df=14
  - t=1-871

- Look up critical value of *t*
  - Use table for two-tailed *t*-test, at *p*=.05, *df*=14
  - critical value = 2.145
  - because *t*=1.871 < 2.145, there is no significant difference
  - therefore, we cannot reject the null hypothesis i.e., there is no difference between the means

df	*.05*	*.01*
1	12.706	63.657
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
6	2.447	3.707
7	2.365	3.499
8	2.306	3.355
9	2.262	3.250
10	2.228	3.169
11	2.201	3.106
12	2.179	3.055
13	2.160	3.012
14	**2.145**	2.977
15	2.131	2.947
…		

Two-tailed Unpaired T-test

Or, use a statistics package (e.g., Excel has simple stats)

Condition one: 3, 4, 4, 4, 5, 5, 5, 6

Condition two: 4, 4, 5, 5, 6, 6, 7, 7

Unpaired t-test

DF:	Unpaired t Value:	Prob. (2-tail):
14	-1.871	.0824

Group:	Count:	Mean:	Std. Dev.:	Std. Error:
one	8	4.5	.926	.327
two	8	5.5	1.195	.423

Common significance tests

Experiment Design	Independent variables (IV)	Conditions for each IV	Types of test
Between-group	1	2	Independent-samples t test
	1	3 or more	One-way ANOVA
	2 or more	2 or more	Factorial ANOVA
Within-group	1	2	Paired-samples t test
	1	3 or more	Repeated measures ANOVA
	2 or more	2 or more	Repeated measures ANOVA
Between- and Within-group	2 or more	2 or more	Split-plot ANOVA

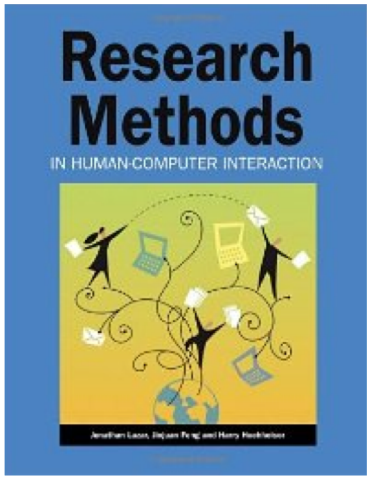
From Lazar et al. Research Methods in Human-Computer Interaction

Remarks on experimental methods

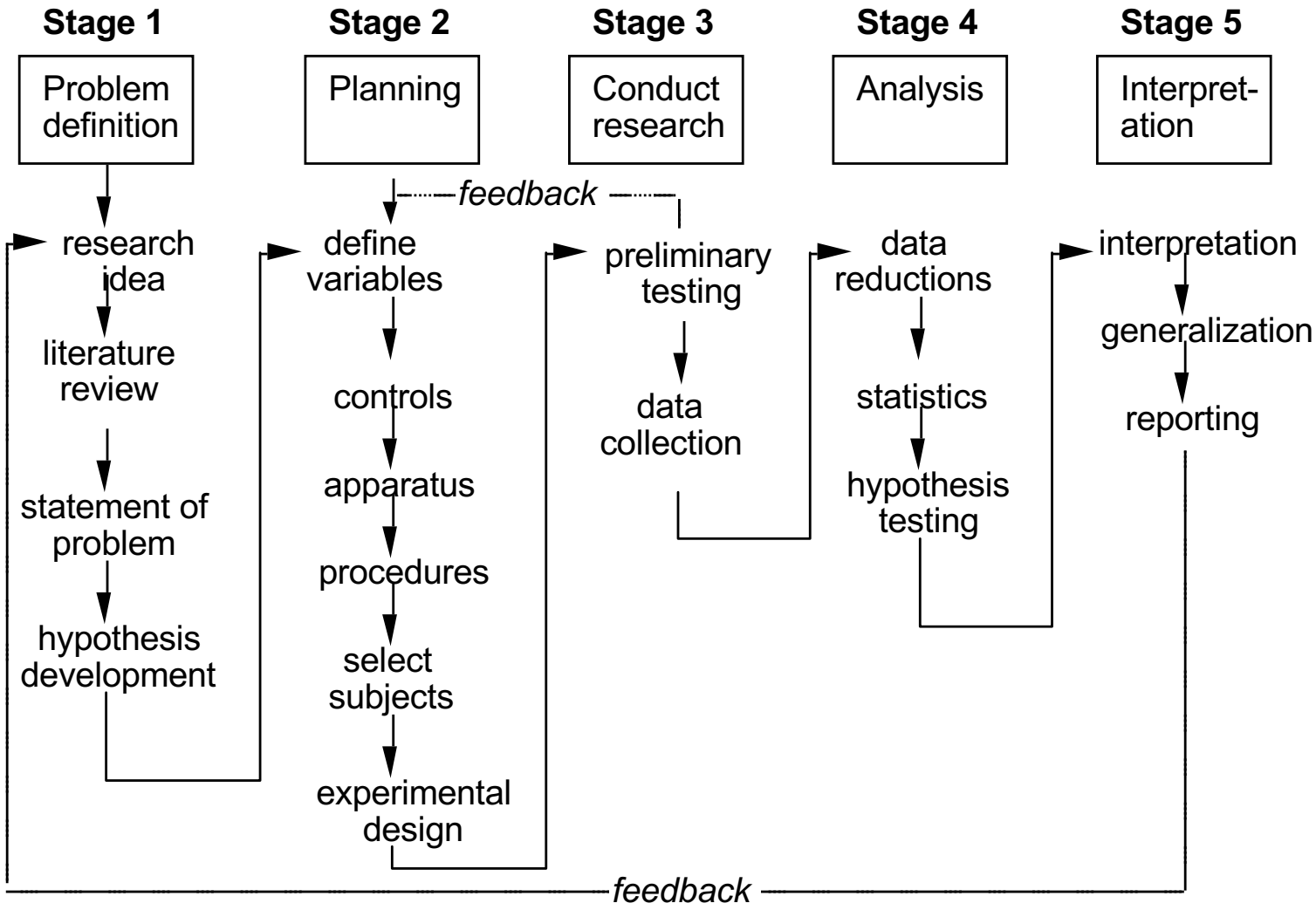
- Remember that t-test and many others (ANOVA) run under strong assumptions over data distribution (normality)
- The experiment can be a lot more complex
  - More levels
  - More independent/dependent variables
  - … but keep it as simple as possible!
- Remember to run a pilot study
  - Test your test
  - Check amount of time required
- How many subjects?
- Specific to HCI
  - Remember that task selection introduces a bias

Tips

- Don’t be obsessed by statistics (and significance)
  - Only necessary condition but not sufficient!
  - Test design and implication of results are a lot more important
- Read ACM CHI papers for real examples to learn from
- Don’t underestimate the task and the effort required
- Be honest with numbers (justify inconsistencies or bad numbers)
- Trust your eyes first
- Use statistical software packages (spss, R, excel)
- Suggested book:
  - **Research Methods in Human-Computer Interaction**, Jonathan Lazar, Jinjuan Heidi Feng, Harry Hochheiser



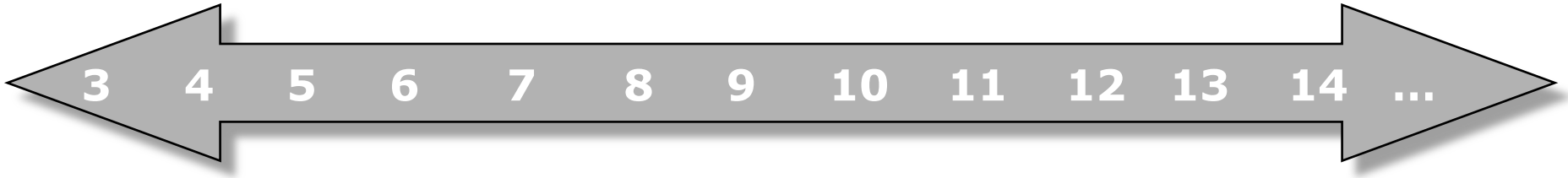
Planning flowchart for experiments



Copied from an early ACM CHI tutorial

How many participants?

- 5-8 participants will find 80% of usability problems (heuristic evaluation)
- At least 8 (or much more depending on your protocol, number of IVs, tasks, etc.) for a controlled experiment.



Qualitative data
Early prototype

Qualitative and quantitative
Later in development

Paired or comparison protocols

Statistically sound results

A/B testing

- Experiment on the web with real users
- Visitors are randomly assigned the solutions A or B (10% of users)
- A cookie is usually assigned, so that individual users always see the same variant.
- Metrics such as click-through rate, success/failure, time are measured for each variant
- The differences are examined for statistical significance

10% of visitors

What you should know

- What is a controlled experiments? What's its purpose?
  - What is a testable hypothesis?
  - What are independent versus dependent variables?
  - How to select subjects?
  - What statistics inform us about?
  - What is the statistical method that can be applied to validate an experimental hypothesis? What does it measure?
- How to design and set up an experiment
- What is A/B testing ?