Multimodal Interfaces

2019

[8] *Evaluating Interactive systems with users*

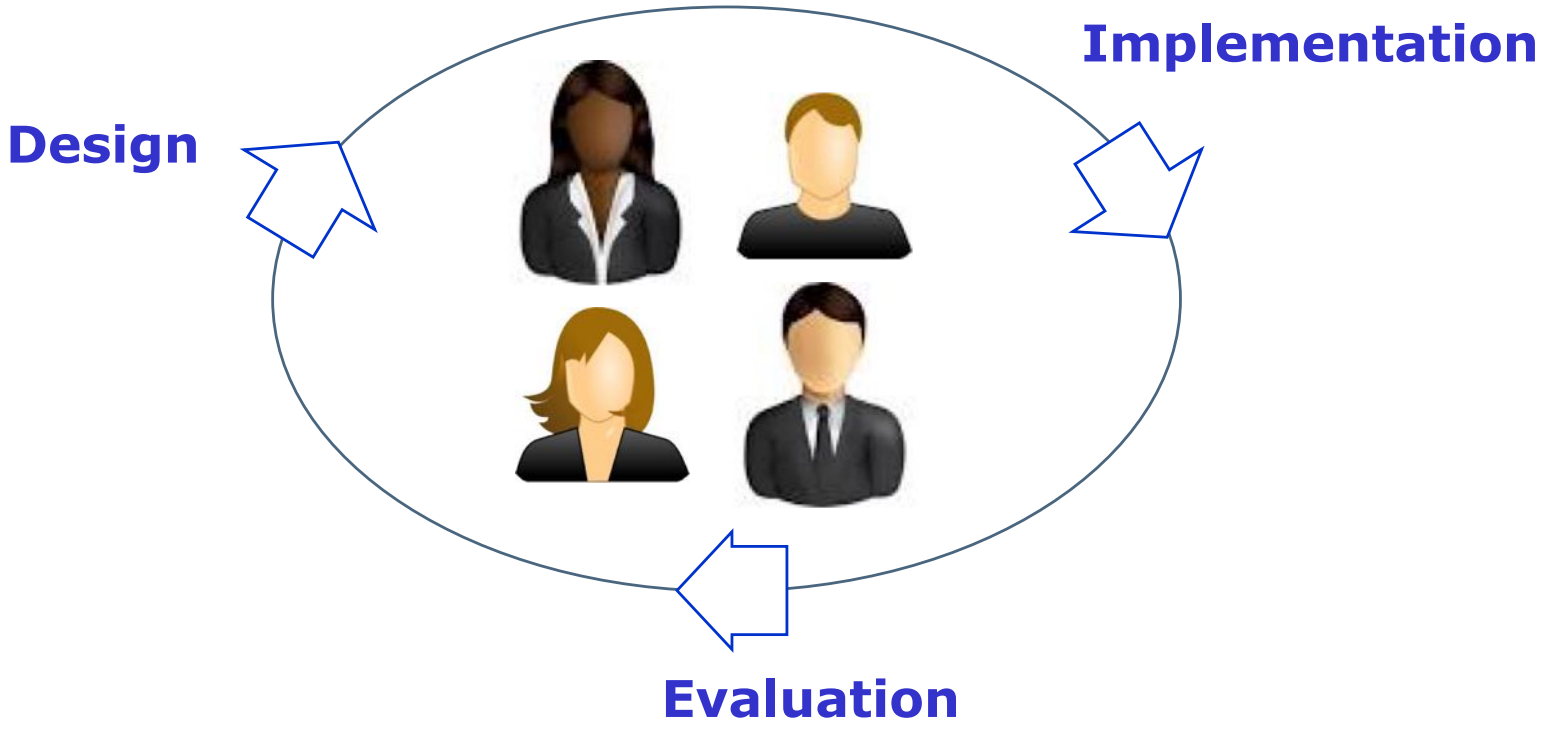
Denis Lalanne

April 9th, 2019

Slides inspired in part from Saul Greenberg HCI class, Lazar book, other sources, and my own thoughts.

Evaluation is not a one-shot process

- Evaluation is part of the development process



…of interactive computing systems for human use

Outline

- Evaluation *with* users:
  - Observation/Testing
  - Questionnaires/Inteviews
  - Controlled experiments
- Goals
  - Gather valuable requirements
  - Observe problems
  - Decide between competing solutions
  - Gain insights over human-interface processes
  - Test theoretical questions
- Evaluation *without* users:
  - Heuristic evaluation -> usability problems

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, requirements
    - ✓ (+) 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
    - ✓ (+) Test theoretical questions, Gain insights over HCI processes
    - ✓ (-) hard methods, difficult to translate in solutions (more about findings)

Qualitative methods (with users)

- Methods
  - direct observation
    - ✓ think-aloud
    - ✓ constructive interaction
  - query techniques (interviews and questionnaires)
  - continuous evaluation (user feedback and field studies)

Direct observations

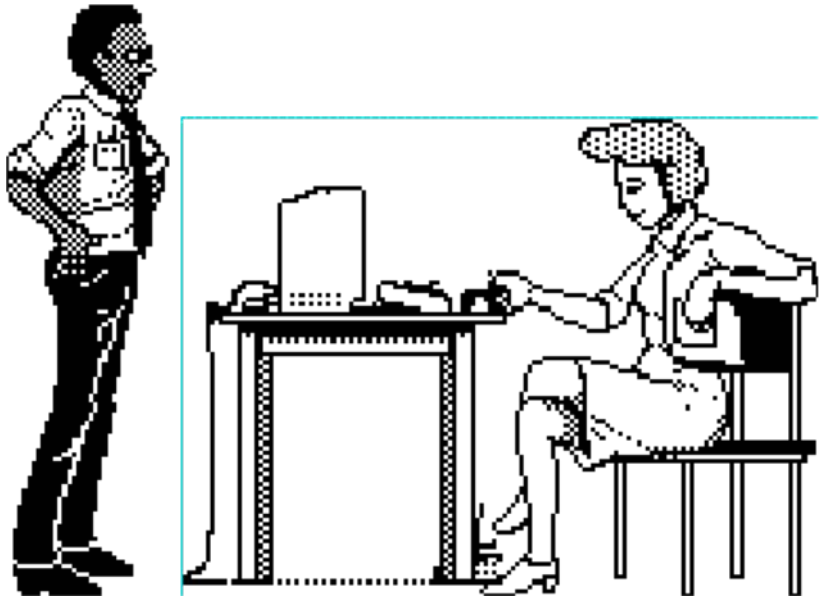
- Evaluator observes users interacting with system
  - in lab:
    - ✓ user asked to complete a set of **pre-determined** tasks
  - in field:
    - ✓ user goes through normal duties

- Value
  - excellent at identifying gross design/interface problems
  - validity depends on how controlled/contrived the situation is

Simple observation method

- User is given the task
- Evaluator just watches the user

- Problem
  - does not give insight into the user's decision process or attitude

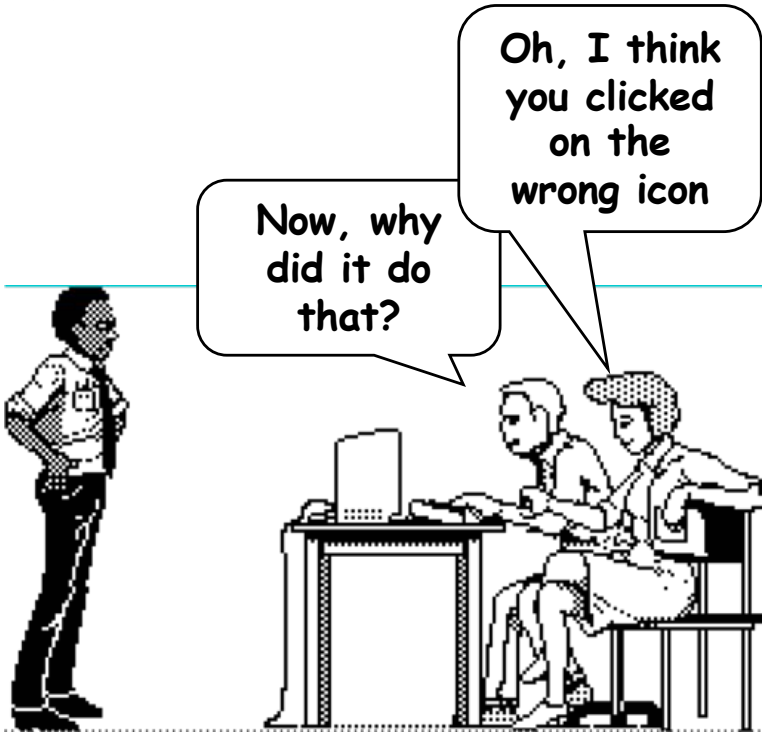


Think aloud method

- Users speak their thoughts while doing the task
  - what they are trying to do
  - why they took an action
  - how they interpret what the system did
  - Pros:
    - ✓ gives insight into what the user is thinking
    - ✓ most widely used evaluation method in industry
  - Cons:
    - ✓ may alter the way users do the task
    - ✓ unnatural (awkward and uncomfortable)
    - ✓ hard to talk if they are concentrating

Constructive interaction method

- Two people work together on a task
  - monitor their normal conversations
  - removes awkwardness of think-aloud
- Co-discovery learning
  - use semi-knowledgeable “coach” and novice
  - only novice uses the interface
    - ✓ novice ask questions
    - ✓ coach responds
  - gives insights into two user groups

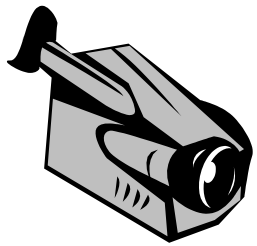


Recording observations

- How do we record user actions for later analysis?
  - otherwise risk forgetting, missing, or misinterpreting events
  - paper and pencil
    - ✓ primitive but cheap
    - ✓ observer records events, comments, and interpretations
    - ✓ hard to get detail (writing is slow)
    - ✓ 2nd observer helps…

  - audio recording
    - ✓ good for recording think aloud talk
    - ✓ hard to tie into on-screen user actions

  - video recording
    - ✓ can see and hear what a user is doing
    - ✓ one camera for screen, rear view mirror useful…
    - ✓ initially intrusive



  - Logging, eye tracking, …

Coding sheet example…

- tracking a person's use of an editor

	General actions			Graph editing			Errors	
Time	text editing	scrolling	image editing	new node	delete node	modify node	correct error	miss error
09:00	X							
09:02				X				
09:05							X	
09:10					X			
09:13								

Questionnaires and Surveys

- Questionnaires / Surveys
  - preparation “expensive,” but administration cheap
    - ✓ can reach a wide subject group (e.g. mail, web)
  - does not require presence of evaluator
  - results can be quantified

- But
  - only as good as the questions asked
- do not ask questions whose answers you will not use!
- determine the audience you want to reach
- determine how would you will deliver / collect the questionnaire
  - web site with forms (e.g. surveymonkey)
  - surface mail

Styles of Questions

- Open-ended questions
  - good for general subjective information
- Closed questions
  - makes questionnaires easy to fill in
  - can be easily analyzed
- Scalar
  - ask user to judge a specific statement on a numeric scale
- Multi-choice
- Ranked
  - useful to indicate a user's preferences
- Combining open-ended and closed questions
  - gets specific response, but allows room for user's opinion

Can you suggest any improvements to the interfaces?

Do you use computers at work:
O often O sometimes O rarely

Characters on the screen are:
hard to read easy to read
1 2 3 4 5

Rank the usefulness of these methods of issuing a command (1 most useful, 2 next most useful..., 0 if not used)
__2__ command line
__1__ menu selection
__3__ control key accelerator

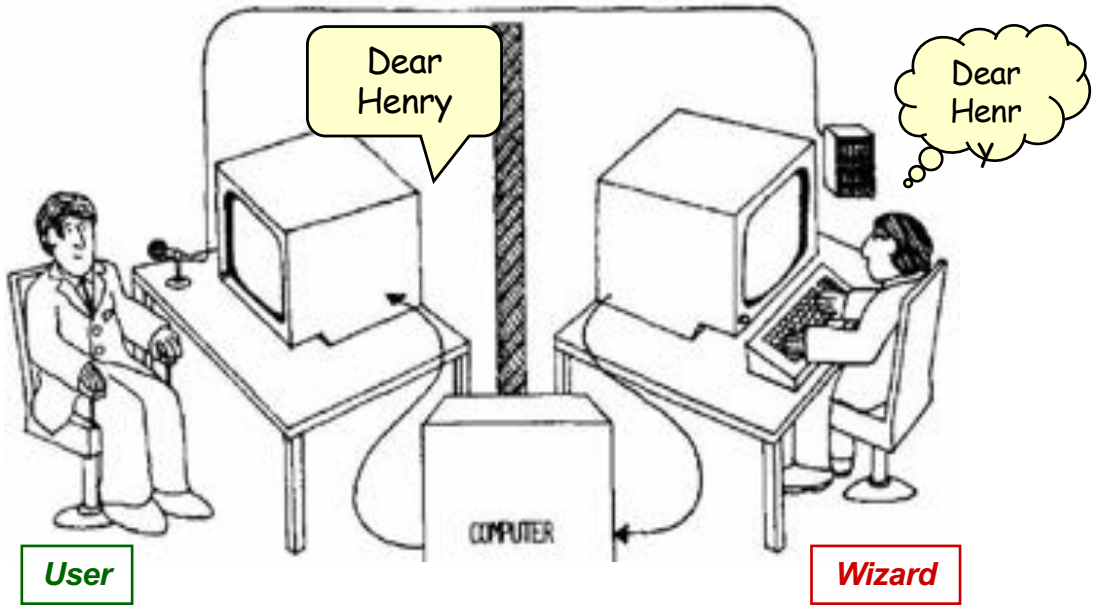
Continuous Evaluation

- Monitor systems in actual use (“in situ”)
  - good for seeing “real life” use
  - usually late stages of development
    - ✓ i.e. beta releases, delivered system
  - fix problems in next release

- User feedback
  - users can provide feedback to designers while using the system
    - ✓ help desks
    - ✓ forums
    - ✓ email
  - best combined with trouble-shooting facility
    - ✓ users always get a response (solution?)

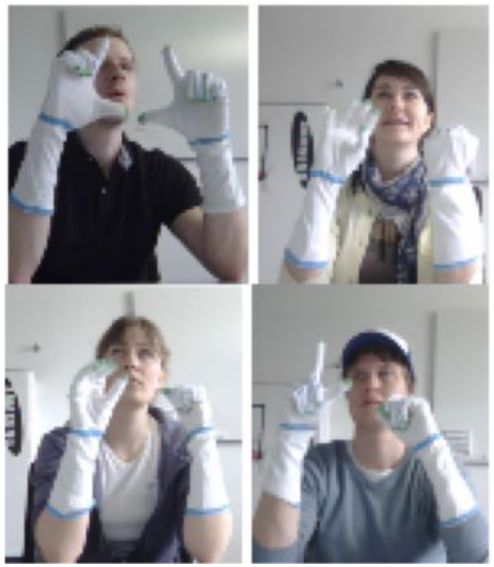
- AB testing

- Possibility to evaluate novel user interface concepts before the technology is mature enough.
- Human 'wizard' simulates system response
  - Interpret user inputs and controls computer to simulate output
- good for:
  - testing "futuristic" ideas



The listening typewriter, IBM 1984

From Gould, Conti & Hovanvecz, Comm ACM 26(4) 1983



WoZ of Gestures
Uni. Fribourg 2010

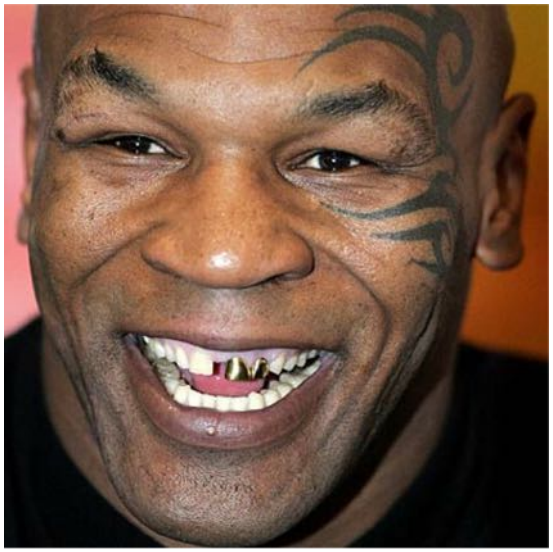
Evaluating interfaces with users

- 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”



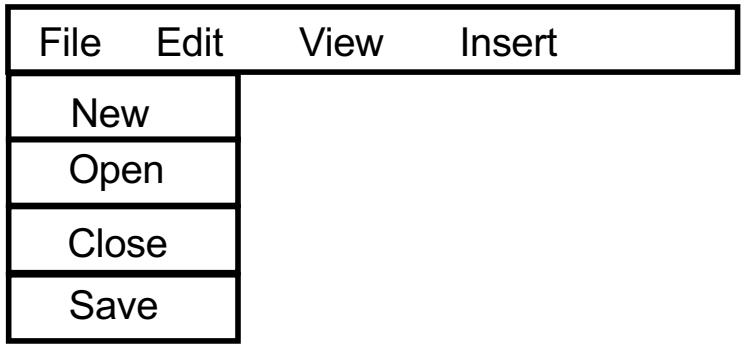
Controlled experiments

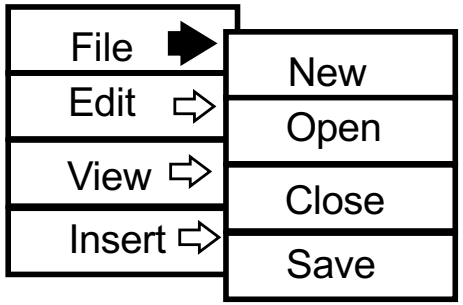
- Phases:
  - A) State a lucid, testable hypothesis
  - Define:
    - ✓ B) Independent variables
    - ✓ C) Dependant variables
  - D) Subject Selection
  - E) Controlling bias
  - F) Statistical analysis
  - G) Interpret your results

A) 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) 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

- 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: 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)
- *in you multimodal interface*
  - ✓ Multimodal commands type: set1 or set2

C) Dependant variables

- 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

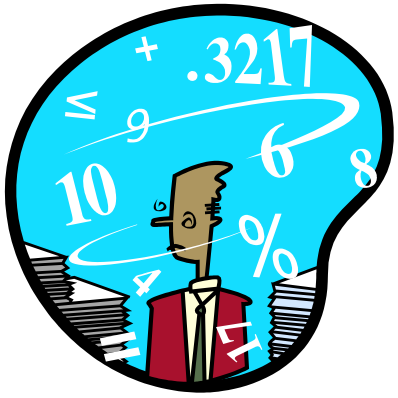
Expert

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
      - – Show tutorial
      - – 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

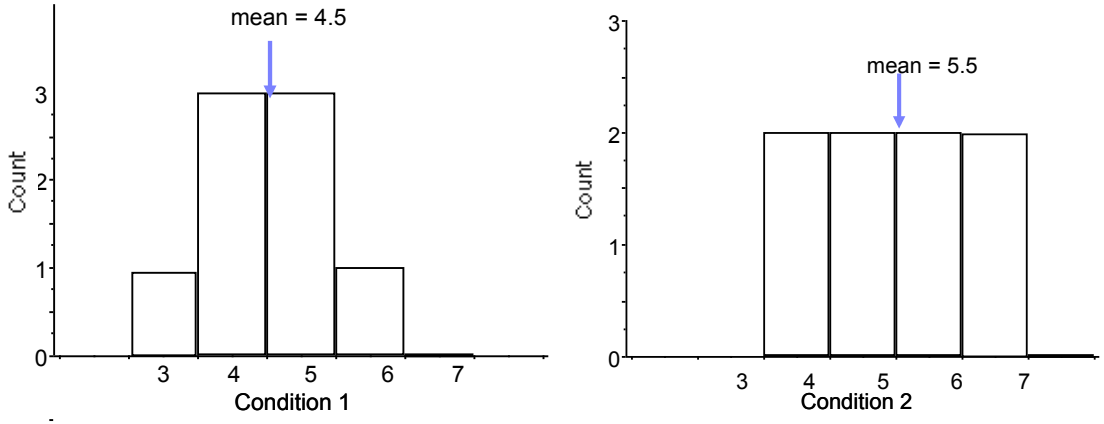
Example: Differences between means

- 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

Example Calculation

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

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

- Calculate t and 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
…		
14	**2.145**	2.977
15	2.131	2.947

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

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
	8	5.5	1.195	.423

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–S10	S1–S10

Condition 2	Condition 1
S10–S20	S10–S20

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

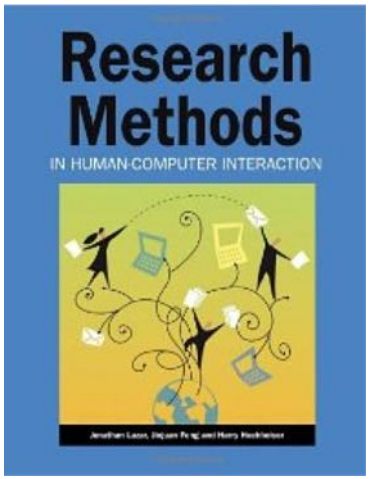
From Lazar et al. Research Methods in Human-Computer Interaction

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! It is very easy to make mistakes otherwise
- 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

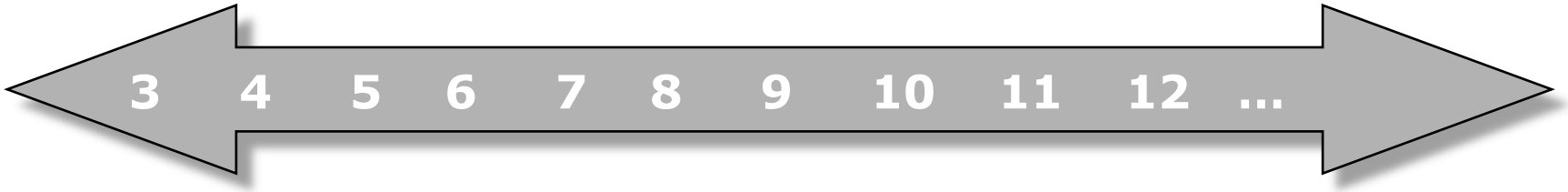
Tips

- 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 CHI or ICMI 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 (even excell with plugins, or R, SPSS)
- Suggested book:
  - **Research Methods in Human-Computer Interaction**,
    - ✓ Jonathan Lazar, Jinjuan Heidi Feng, Harry Hochheiser



How many participants?

- 5-8 participants will find 80% of usability problems (heuristic evaluation)
- At least 8 (or more depending on your protocol) for a controlled experiment to find statically significant differences



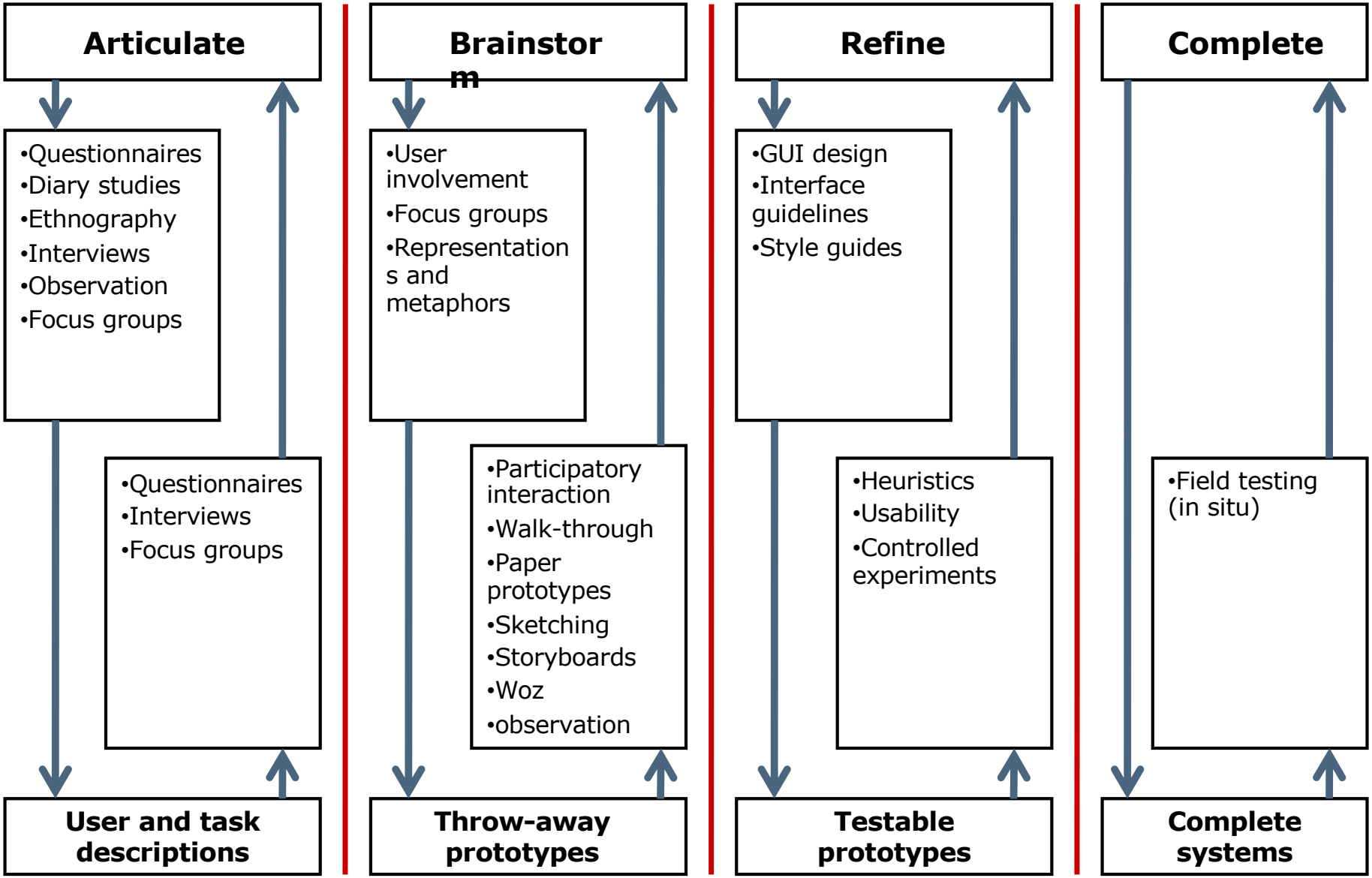
Qualitative data
Early prototype

Qualitative and quantitative
Later in development

Paired or comparison protocols

Statistically sound results

UCD process

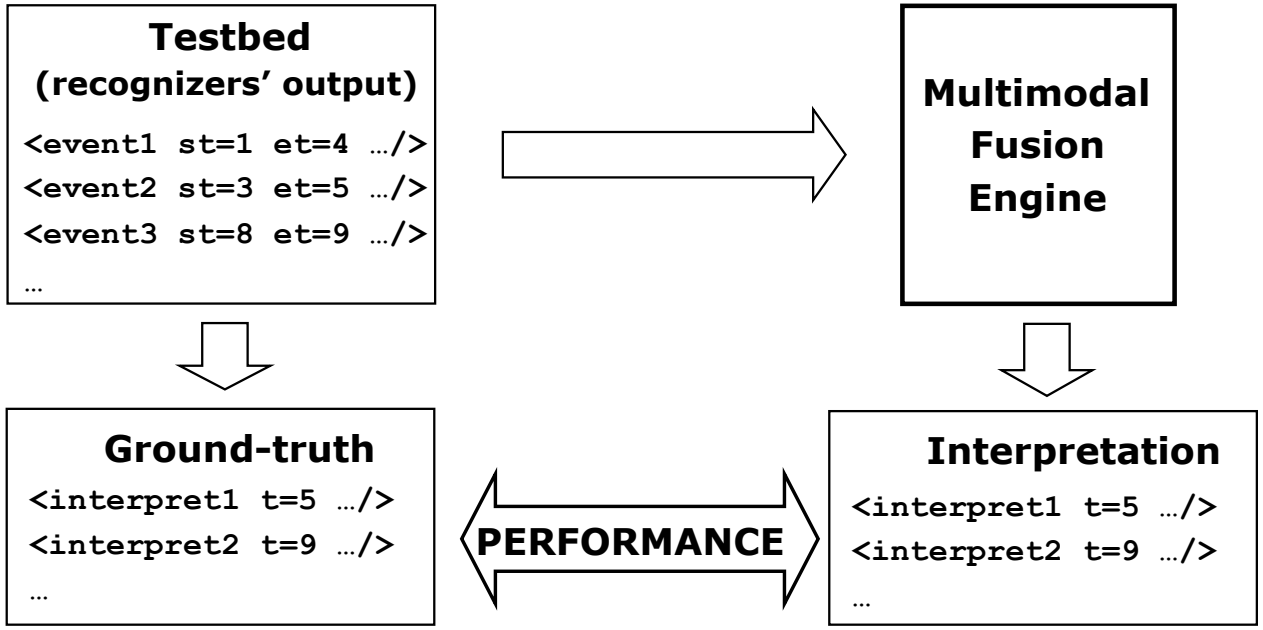

Articulate
•Questionnaires
•Diary studies
•Ethnography
•Interviews
•Observation
•Focus groups
•Questionnaires
•Interviews
•Focus groups
User and task descriptions
Brainstorm
•User involvement
•Focus groups
•Representations and metaphors
•Participatory interaction
•Walk-through
•Paper prototypes
•Sketching
•Storyboards
•Woz
•observation
Throw-away prototypes
Refine
•GUI design
•Interface guidelines
•Style guides
•Heuristics
•Usability
•Controlled experiments
Testable prototypes
Complete
•Field testing (in situ)
Complete systems

Multimodal Interactions

- First use the CASE/CARE to characterize your interactions
- Typical questions
  - Which modality (or combination of modalities) do users prefer?
  - With which modality (or combination of modalities) are they more efficient?
  - Is there a good adequation modality (or combination of modalities) with the task?
- Characterize the types of errors
  - User errors
    - ✓ E.g. user says a word or make a gesture that does not exist
  - Recognition errors
    - ✓ the word pronounced correctly was not recognized by the system
  - Fusion errors (interpretation)
    - ✓ the modalities were correctly recognized independently but the fusion is incorrect (time synchronicity problem, interpretation error)

Multimodal fusion performance

- You might be interested to quantitatively measure the following:
  - **Response time**: time the multimodal system takes to return an interpretation after receiving multimodal inputs.
  - **Confidence**: Confidence of machine response, based on confidence scores in the testbed
  - **Efficiency**: success or failure of the multimodal system to interpret correctly the testbed entries.



What you should be able to answer by now

- What are qualitative versus quantitative user evaluation methods? What are their respective goals?
- What are the qualitative user evaluation methods?
- Why controlled experiments can provide clear convincing result on specific issues?
  - What is a testable hypothesis?
  - What are independent versus dependent variables?
  - How to select subjects?
  - What statistics inform us about? What are the available methods?
- What are the particularities associated with the evaluation of multimodal systems?