Evaluation of User Interfaces

Contents

- Introduction
- The User in the Loop
- Interacting with virtual worlds
- Immersion, Presence and Embodiment
- Evaluation of user interfaces
  - Concepts and definitions
  - Evaluation tools
  - Evaluation methods
  - Evaluation metrics
  - Evaluation methodology
  - Challenges for 3DUI/VR evaluations

Usability

- The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.
- Usability is a **non-functional** requirement. As with other non-functional requirements, usability cannot be directly measured but must be quantified by means of indirect measures or attributes such as, for example, the number of reported problems with ease-of-use of a system.

System Acceptability

Quantitative

- **Efficiency**: How quickly can user perform tasks after they are trained?
  - Users' goals are realized
  - User tasks done better, easier, or faster
- **Errors**: How many errors do users make, how severe are these errors, and how easily can they recover?

Qualitative

- **Learnability**: How easy is to accomplish basic tasks the first time they use the system?
- **Memorability**: How easily can they re-establish proficiency after a period of not using the system.
- **Satisfaction**: How pleasant is it to use the design?
  - Users are not frustrated
  - Users are not uncomfortable

Purposes of Evaluation

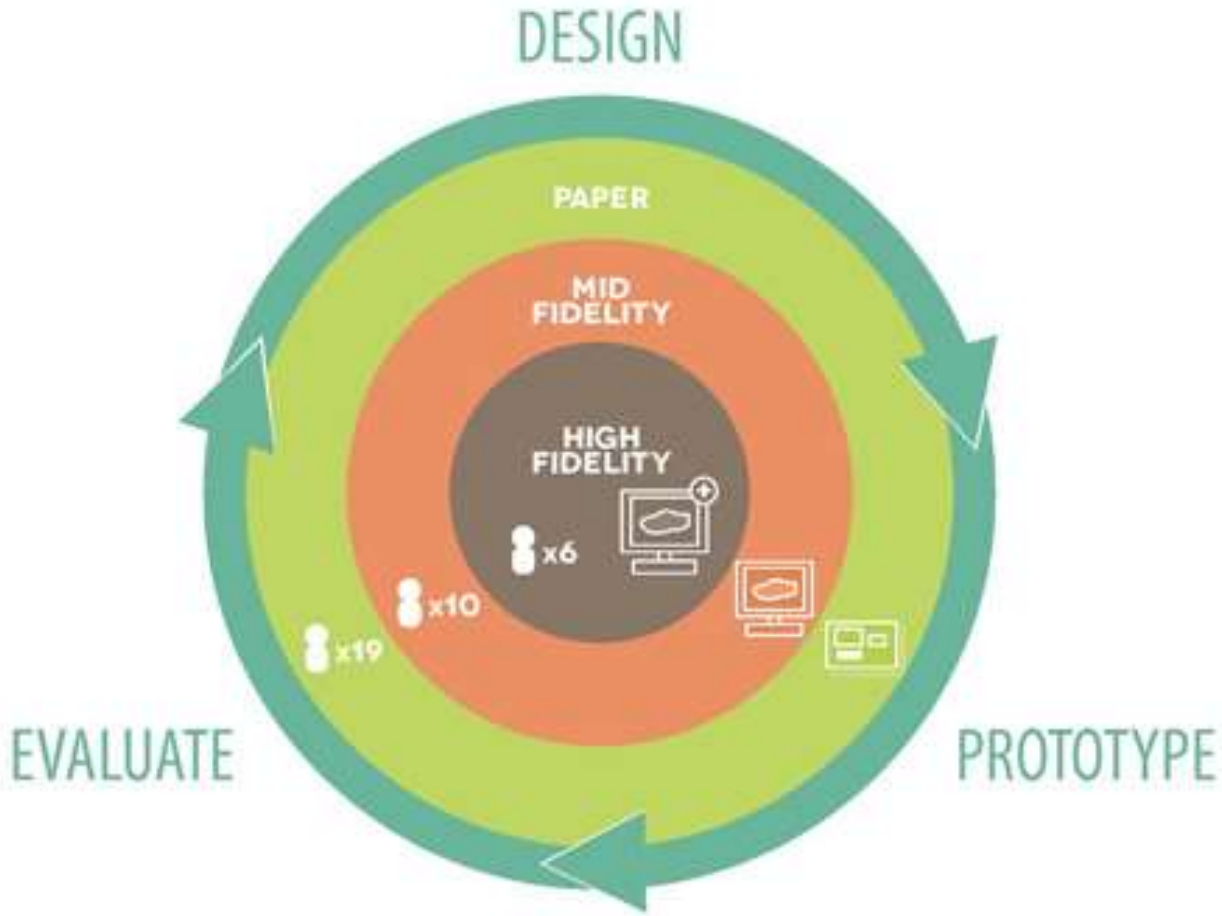
- Analysis, **assessment** and testing of a component or technique.

- Identification of usability **problems**
  - Performed in an iterative fashion.
  - Each user performs in a different way.
  - **Critical** step in any system.

- Acquire general understanding of the **usability**
  - **Knowledge** about design comes from evaluation.
  - Creation of design **guidelines**.

- Develop **performance** models
  - Predict performance on a particular task (e.g. Fitts' Law).

Purposes of Evaluation

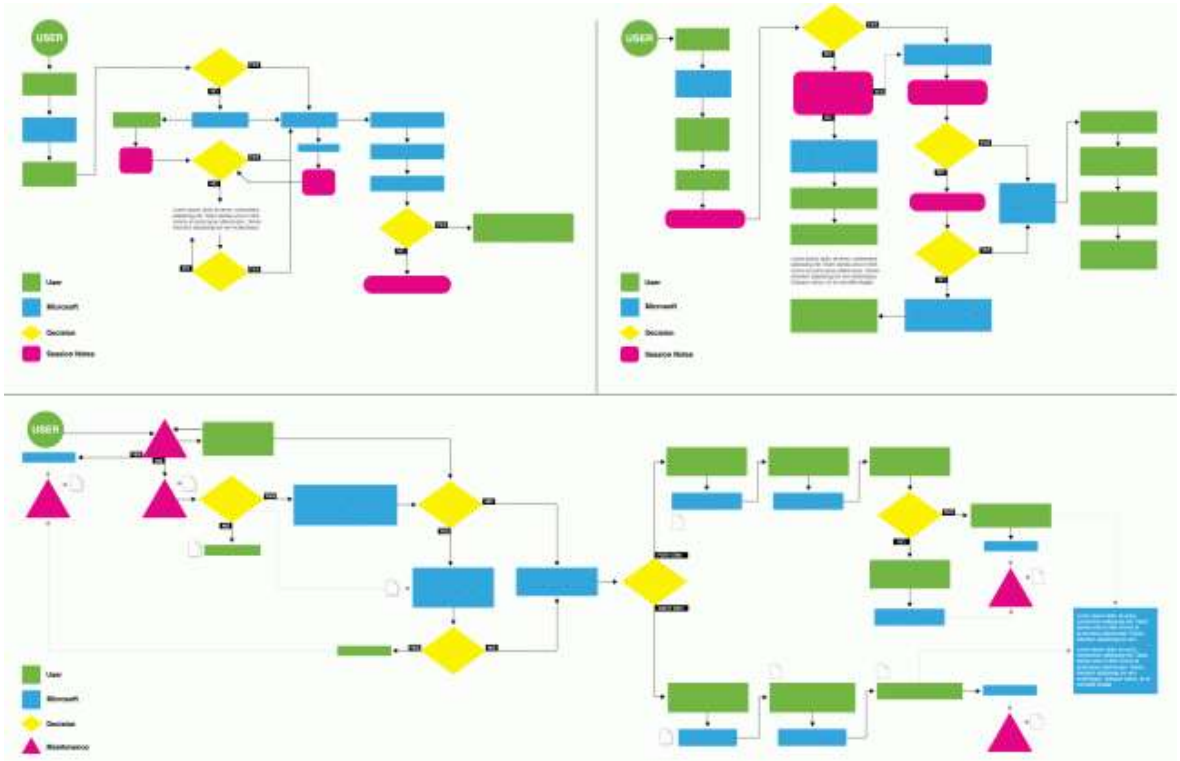


User Interface Evaluation

- Concepts and definitions
- Evaluation tools
- Evaluation methods
- Evaluation metrics
- Evaluation methodology
- Challenges for 3DUI/VR evaluations

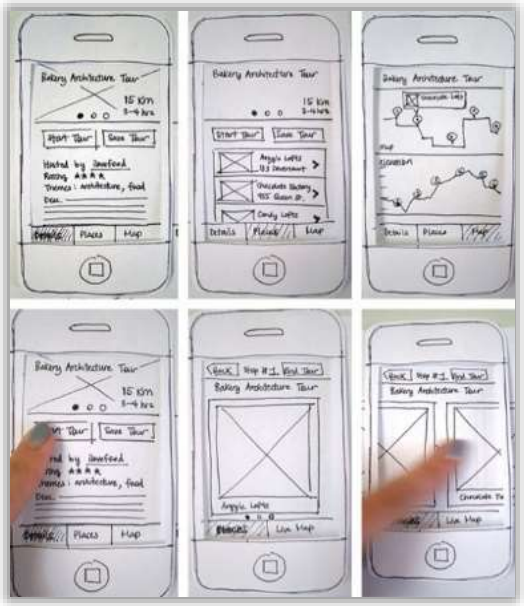
Task Analysis

- User **task** analysis
  - Determine **what** users will do with the application
  - Based on extensive input from representative users



Prototyping (Not always simple in 3D…)

- Usability evaluation can be obtained from low-fidelity **prototypes**
  - E.g. Paper-based, static mockups
  - Should not be required to be complete
  - Should be easy to change
- Strategy for efficiently dealing with things that are hard to predict



Wizard of Oz

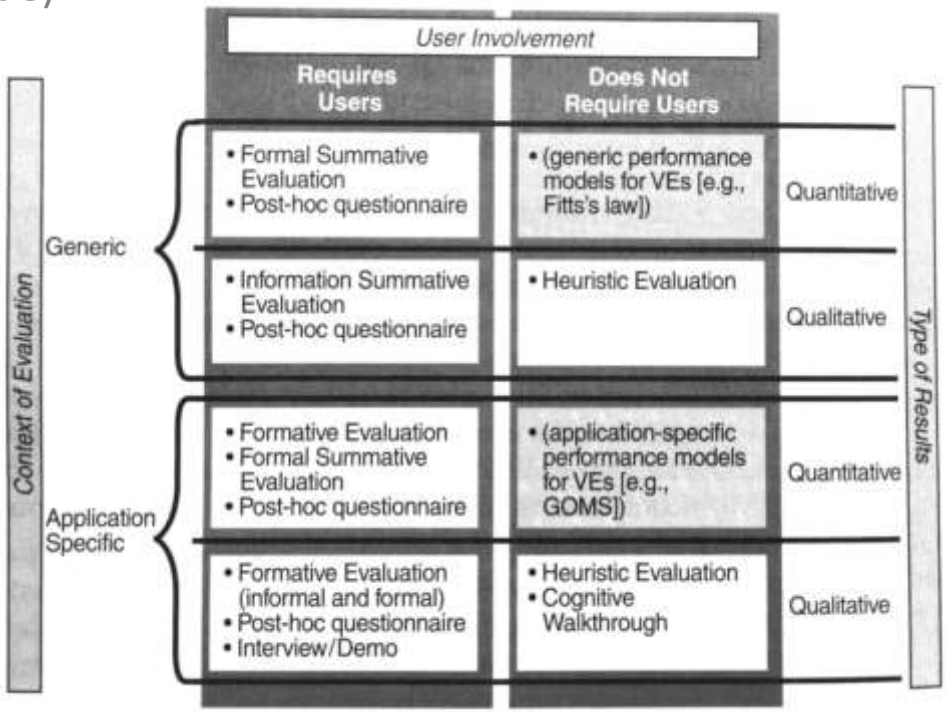
- A human provides the functionality missing in the prototype
  - "Simulates the behavior of a theoretical intelligent computer application"
- Test systems that present implementation challenges
  - Allows for testing the usability of the system before its development

User Interface Evaluation

- Concepts and definitions
- Evaluation tools
- Evaluation methods
- Evaluation metrics
- Evaluation methodology
- Challenges for 3DUI/VR evaluations

Evaluation Methods

- User Involvement
  - Requires (final) users
  - Does not requires users
- Content of the evaluation
  - Generic (e.g. interaction technique)
  - Application specific
- Type of the results
  - Quantitative
  - Qualitative



Heuristic Evaluation

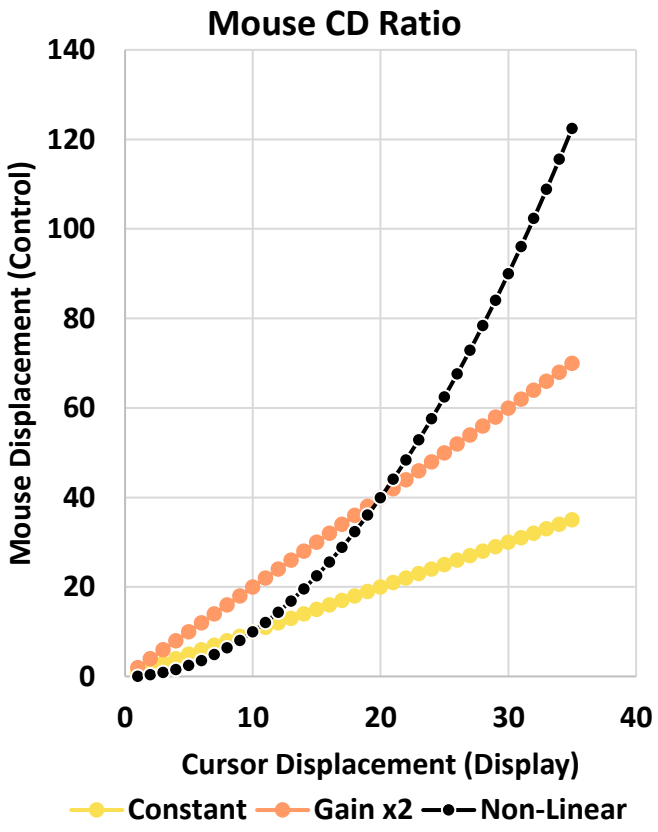
- Used in the **early steps** of the design
- Performed by usability **experts**
  - No real users involved in the evaluation
- Based on **guidelines** and heuristics
  - Find common flaws
  - Qualitative evaluation

Formative Evaluation

- Based on observational and **empirical** evaluation.
- **Users** have to perform a set of tasks.
  - Quantitative and qualitative **data** is gathered.

Sumative Evaluation

- Aims at **comparing** two or more UI designs
- Several versions of the interfaces are tested
  - Input devices
  - Interaction techniques
  - …



User Interface Evaluation

- Introduction
- Evaluation tools
- Evaluation methods
- Evaluation metrics
- Evaluation methodology
- Challenges for 3DUI/VR evaluations

Task Performance

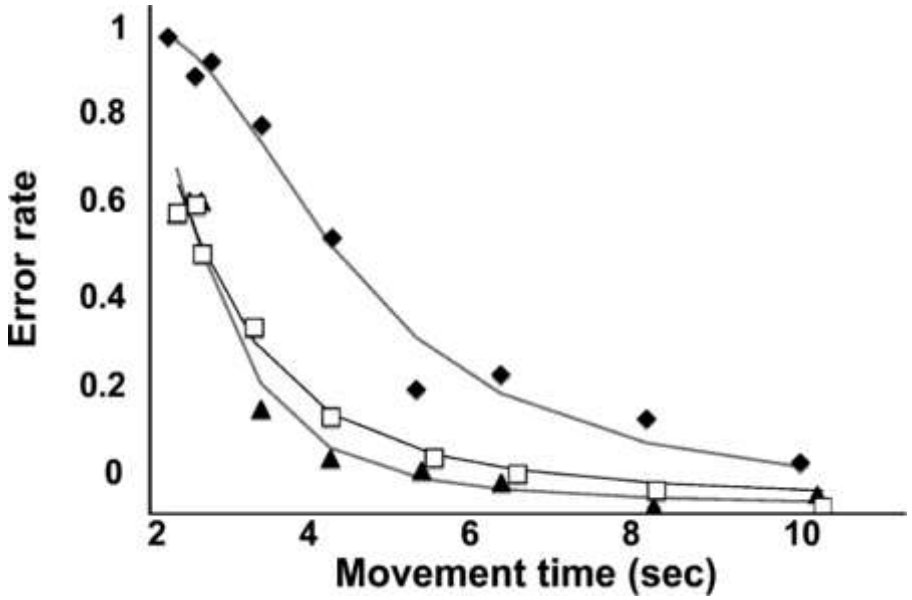
- Direct measure of the user performance

- Objective Measures
  - Task completion time
  - Number of errors
  - Accuracy / Precision

- Domain-specific metrics
  - Education: learning
  - Training: spatial awareness
  - Design: expressiveness (evaluated by experts)

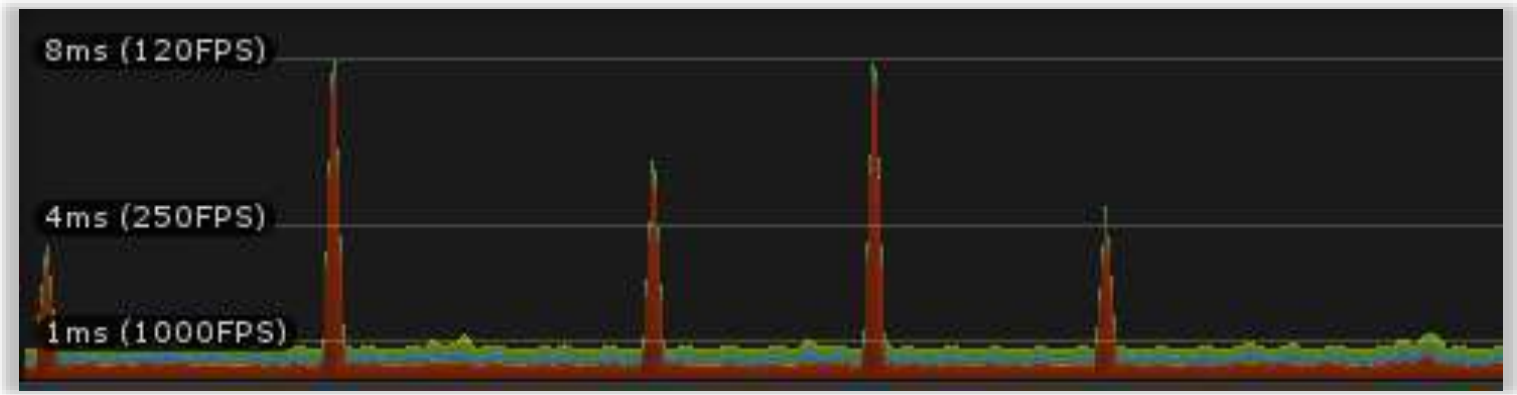
Task Performance

- Most important metrics : **time** and **accuracy**
  - Tradeoff between speed and accuracy
- Evaluation strategies
  - As quick as precise as possible (participant decision)
  - As quick as possible
  - As precise as possible



System Performance

- Need to assess the performance of the system
  - End-to-end latency
  - Frame rate (average, jitter, minimum)
  - Network delays
  - Recognition accuracy
- Critical issue unless the user's experience is not altered



User Preferences / Subjective

- Subjective perception of the user interface
  - Ease of use
  - Ease of learning
  - Satisfaction
  - Suggestions of improvement
- User comfort
  - Simulator sickness
  - Physical fatigue (arms/hands/eyes)
- Verbal protocol taking
  - Participants think aloud, talking while performing tasks
  - Can be intrusive, but effective
  - Some participants not good at talking

Information Gathering

Pre

- Demographic data
  - Can explain biases in quantitative and qualitative data
  - E.g. age, gender, experience, eye dominance, hand dominance.
- Relevant user mental/cognitive abilities
  - E.g. spatial reasoning tests

During

- Data logging
  - Log as much as you can!
  - *Cook* your data as much as possible -> speed up the analysis

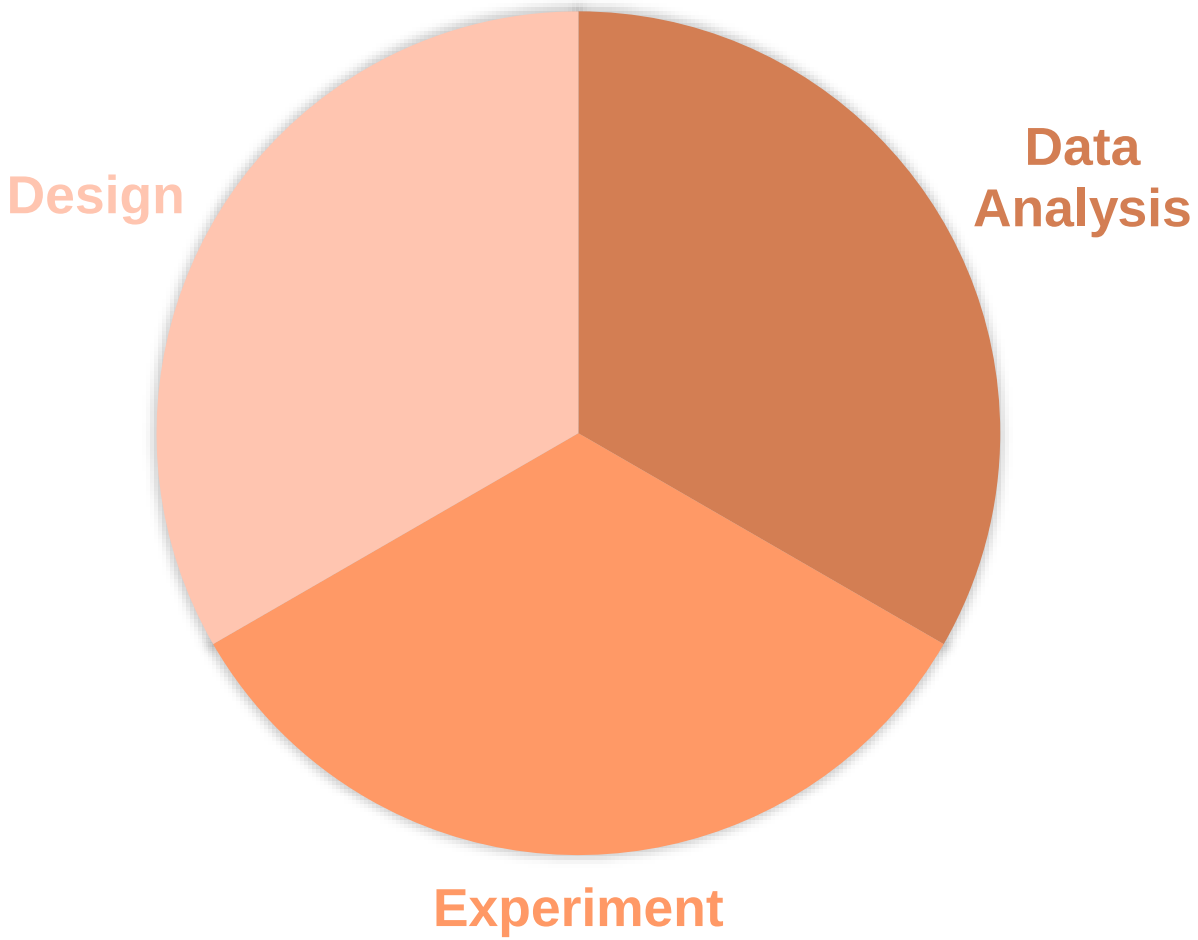
Post

- Subjective Questionnaires
  - Rate a written set of questions
  - Open questions
- User Interviews
  - Obtain information directly from users
  - Structured or open-ended

User Interface Evaluation

- Introduction
- Evaluation tools
- Evaluation methods for 3D UI
- Evaluation metrics
- Evaluation methodology
- Challenges for 3DUI/VR evaluations

Major steps in user experimentation



Experimental Design

- Define the main theory to evaluate
- Define experimental hypothesis in order to validate the main theory
  - Define dependent (measures) and independent variables (factors)
  - Define the analysis that will be required to test the experimental hypothesis
- Define the experimental protocol
  - Presentation order (factors)
  - Revisit the design if the experimental is too long!!!
- Develop the system used in the experiment
  - Presentation, instructions, logging, conditions,…
- Pilot testing
  - Ensure that users understand the protocol
  - Ensure that data was logged correctly

Independent Variables

- They might have an impact on the outcome on the experiment
- They are not correlated with other variables
- Manipulated by the experimenter (experimental conditions)

- Example: Visual appearance of the user's avatar
  - Three levels: sphere, robot, real
  - Encode the realism of the representation

Independent Variables

- Within-subjects variable
  - Each participant will be exposed to all levels of the independent variable
  - Requires less subjects
  - Users can subjectively compare the conditions (if relevant)
  - Statistical tests will be able to find smaller effects
  - Need to minimize the potential ordering effects
    - Random ordering (if the number of combinations is unmanageable)
    - Latin Square design
    - Counterbalancing (all possible combinations)
    - The ordering will determine the number of participants (e.g. multiple of 4)

$$[1] \quad \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \\ 3 & 4 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \\ 3 & 1 & 4 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}$$

Latin-Square design

Independent Variables

- Between-subjects variable
  - Each participant will be exposed to one level of the independent variable
  - Avoid order effects between different levels.
  - Decrease the length of the experiment
  - Requires the increase of the sample size (participants)
  - It can be problematic in highly heterogeneous populations
  - Example: Control group

Animated

Rigid

Independent Variables

- Mixed designs
  - Mix between and within-subjects variables
  - Can complexity the statistical analysis
  - Grouping and counterbalancing should be handled with care

Dependent Variables

- Objective Measures
  - Task completion time
  - Number of errors
  - Accuracy / Precision
- Domain-specific metrics
  - Learning, spatial awareness, expressiveness
- Subjective responses
  - Ease of use, ease of learning, satisfaction
  - Simulator sickness
  - Physical fatigue (arms/hands/eyes)

Experimental Platform

- Implement the required features
  - Interaction techniques, virtual environment, …
- Platform requiring a minimal interaction from the experimenter
  - Avoid errors and reduce bias
  - Minimize oral instructions from the experimenter
- Automatic data logging
  - Ensure that all data is recorder in the same conditions
  - Cook data as much as possible
  - Anonymize the data
- Test your platform
  - Pilot testing, ask advice / suggestions…

Beware of the Clever Hans Effect!

- A horse was claimed to have been able to perform arithmetic tasks.
  - The horse was just responding to the reaction of the observers.
  - Generation of an observer-expectancy effect

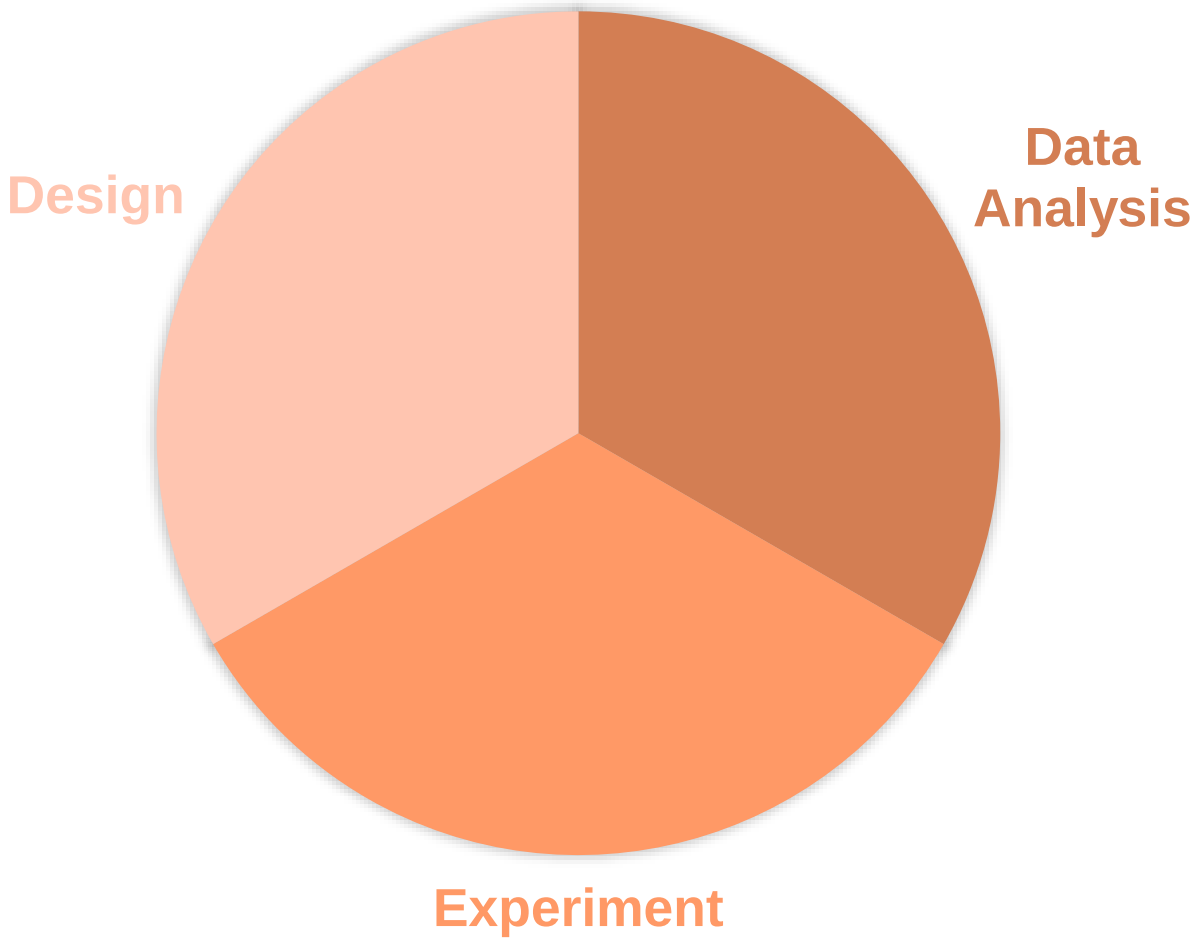
- Clever Hans effects are likely to occur in experiments with humans.
  - Bias of the experimented due to preconceived hypothesis
  - Use double-blind protocols: neither the experimenter nor the subject knows what condition the subject is in, and thus what his or her responses are predicted to be.
  - Replacing the experimenter with a computer which provides standardized instructions and logging without giving clues.

https://en.wikipedia.org/wiki/Clever_Hans

Ethical Regulations

- Respect the Helsinki declaration (1964)
  - Ethical principles for the research with human subjects
    - Obligation to inform about the nature, the protocol, the risks and they right to stop the experiment whenever they want.
    - Confidentiality of their data
- In practice
  - Experiments have to be approved by an ethical committee (Institutional Review Board)
    - *Comités de protection des personnes* (CPP)
    - Local ethical committees
    - Needed for certain journals and conferences
  - Write a informed consent form and the description of the experimental protocol which will be signed for each participant
  - Ensure the data anonymization

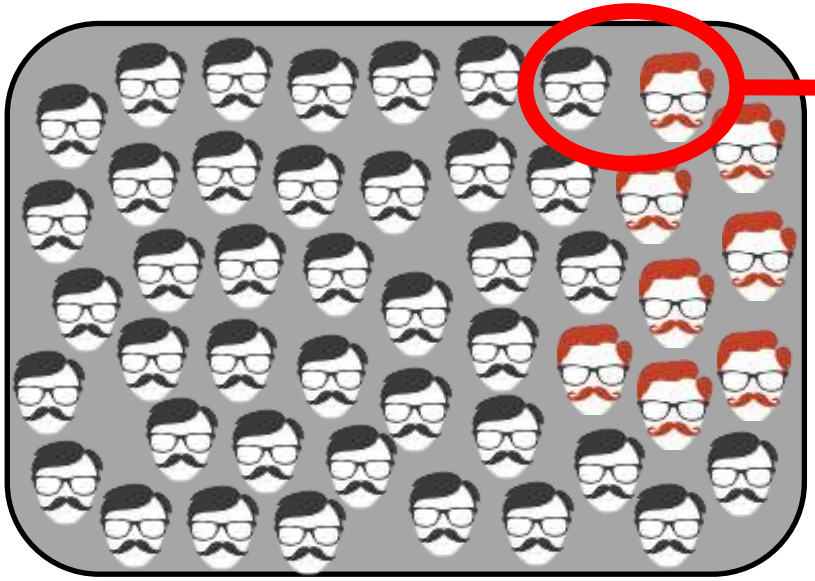
Major steps in user experimentation



Recruiting the participants

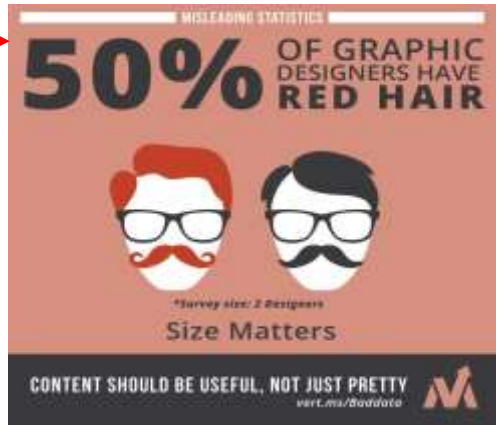
- Recruit participants ensuring a representative sample

Population



Graphic designers

Sample

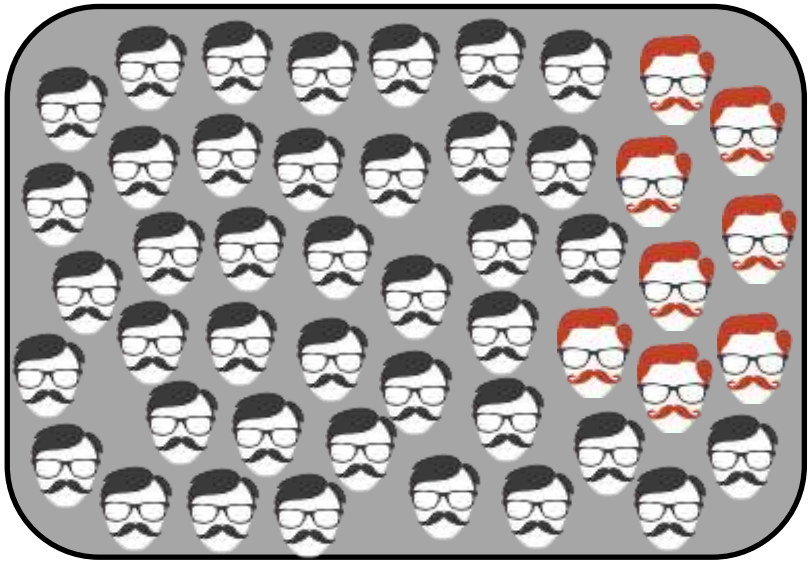


The sample must be representative

Recruiting the participants

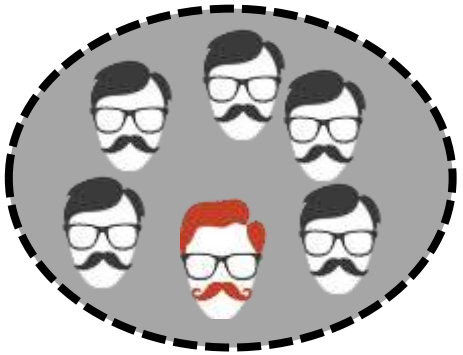
- Recruit participants ensuring a representative sample

Population



Graphic designers

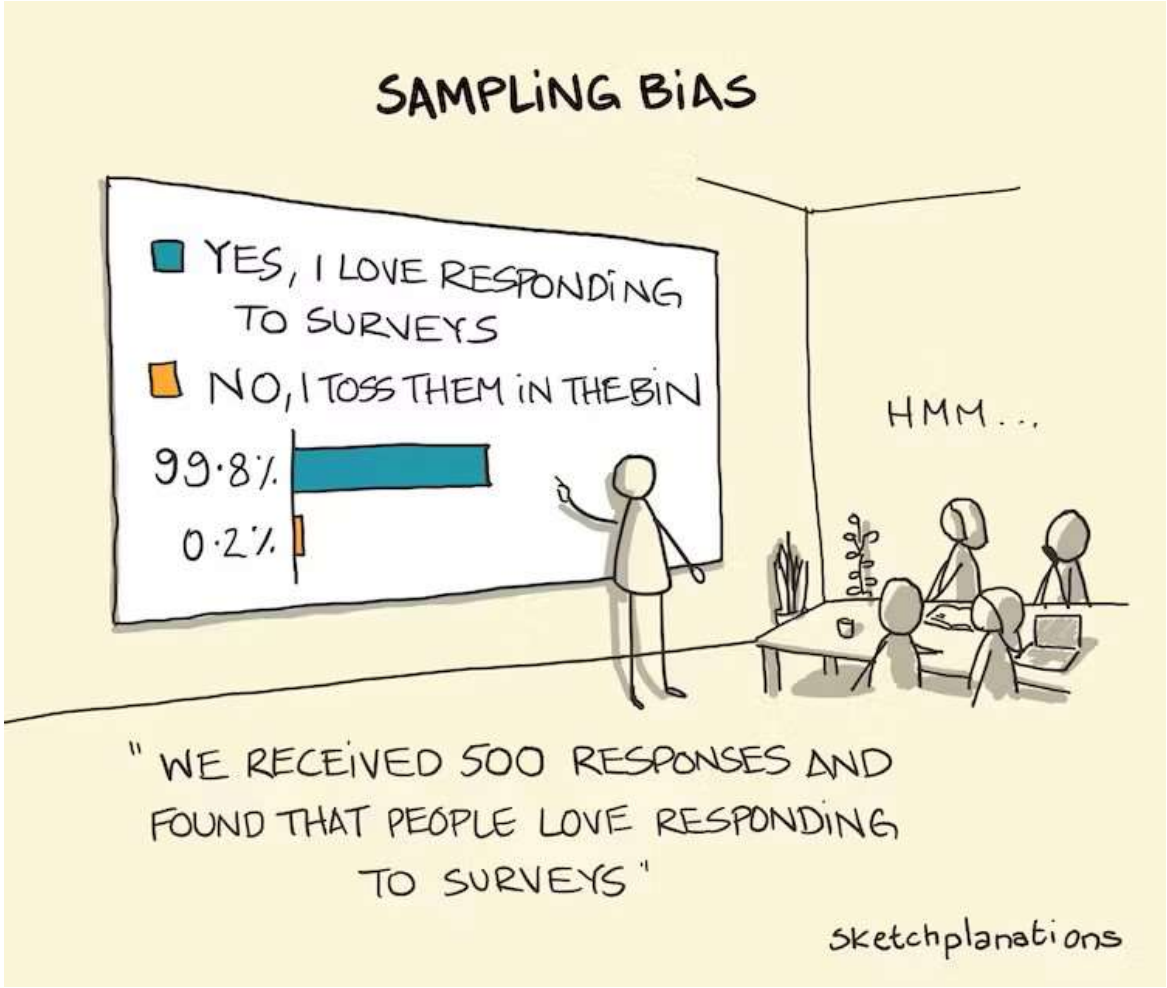
Sample



The sample must be representative

Recruiting the participants

Recruiting the participants



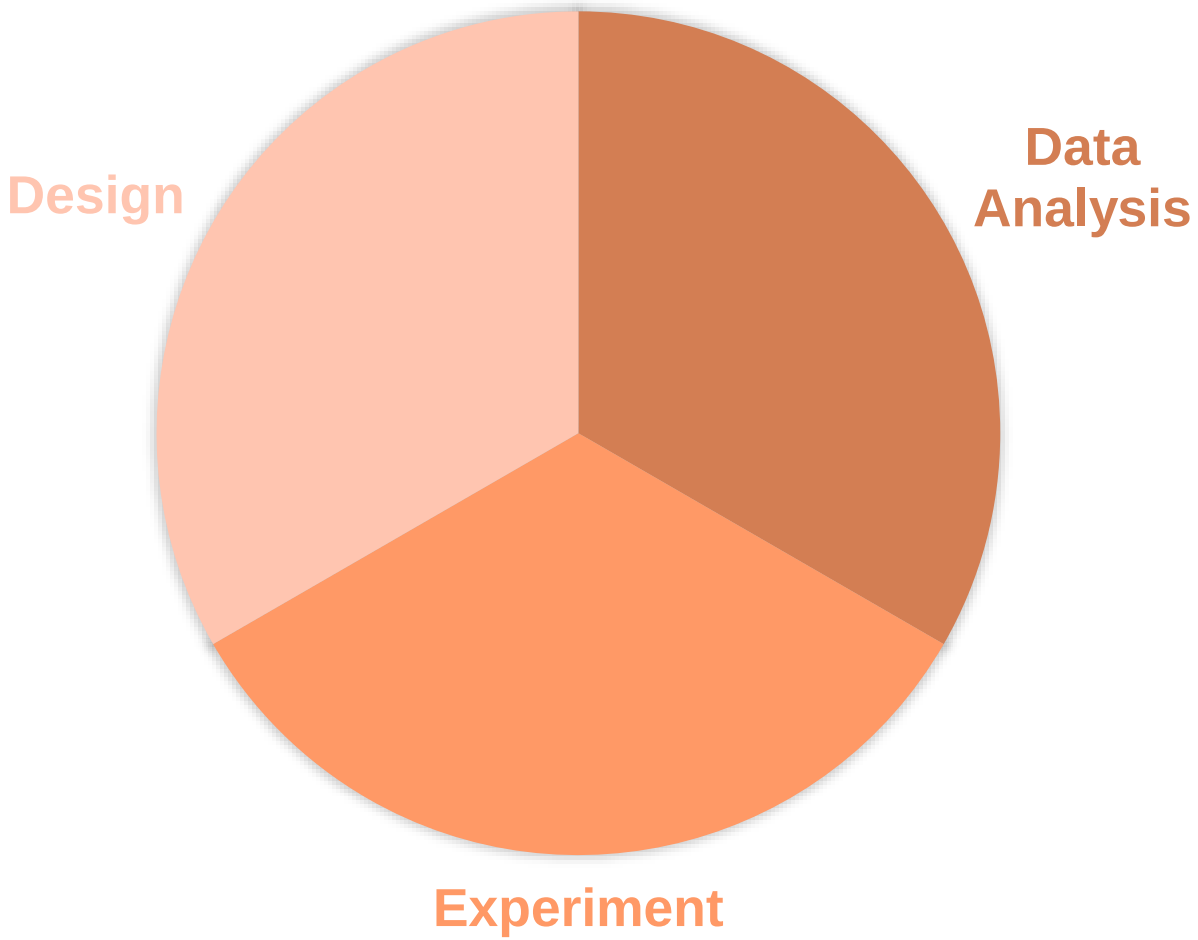
Preliminaries with participants

- Explain protocol to participant, including any compensation
  - This can be written in the informed consent form
- Have participant sign informed consent form (and NDA)
  - Explain the rights and duties of the participant
  - Ask for consent if you take pictures or videos
- Show participant the experimental set-up if they are interested

During the experiment

- Follow the plan defined by the experimental protocol
- Ensure that everything goes as expected
  - Monitor the actions of the user
  - Avoid the an Observer effect!!
- **Critical incident**: something that happens during the experiment that might have a significant effect on the results
  - Responsibility of the experimenter to identify and record critical incidents
  - Critical incidents are indicators of usability problems
  - Very important evaluation data!
  - Later analyze the problem and cause within the interaction design
- **Avoid any change on the experimental protocol**
  - **It will require to throw away all gathered data!!!**
  - **Always hard to justify in a paper….**

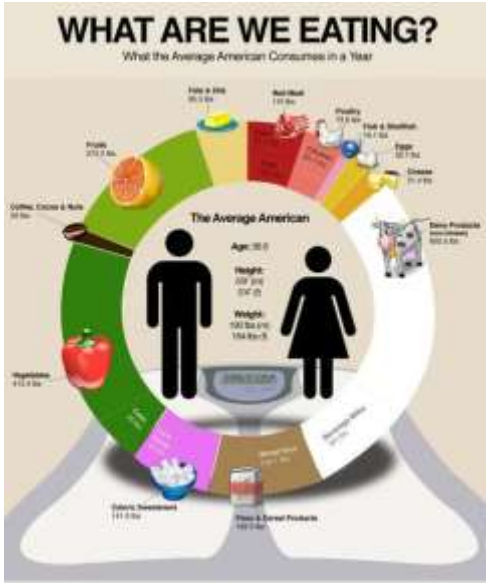
Major steps in user experimentation

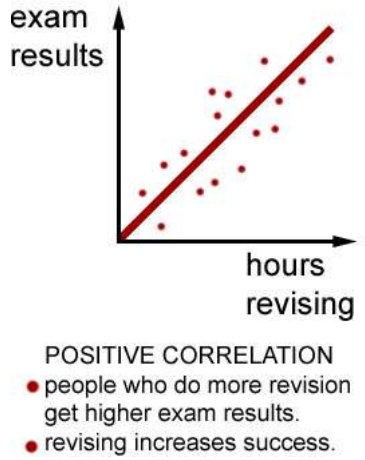


Statistics!

- Descriptive Statistics
  - Transform and summarize
  - E.g. Mean, Median, plots

- Inferential Statistics
  - Generalize the results to the entire population. Hypothesis testing.
  - E.g. Evaluate the relation between variables

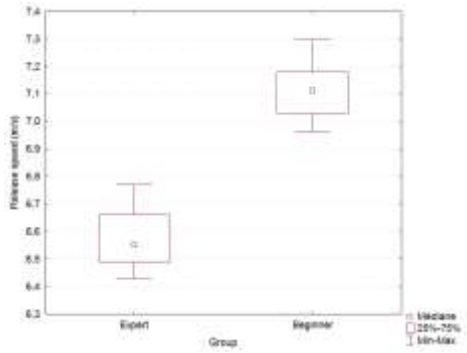


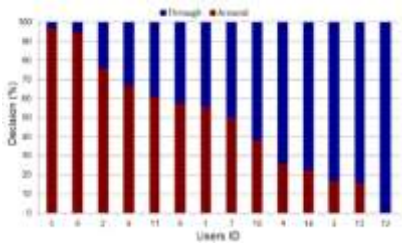


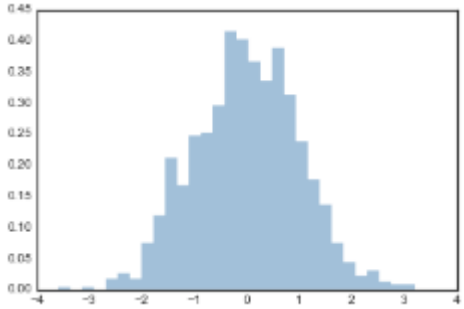
R^2=0,92
p<005

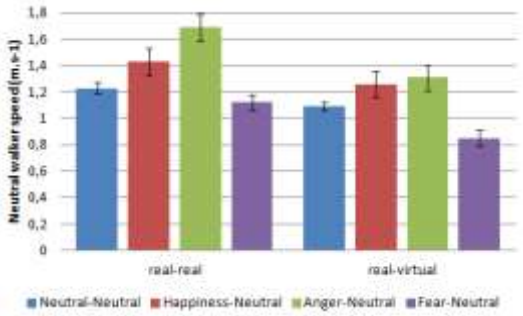
Visualize your data

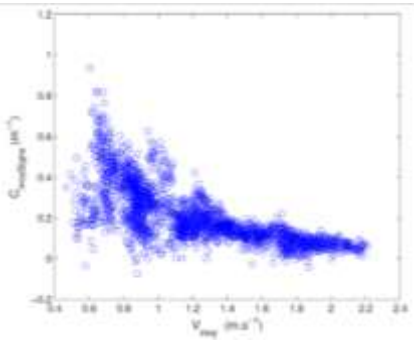
- The dataset structure isn't just a summary
  - Explore the data to find patterns



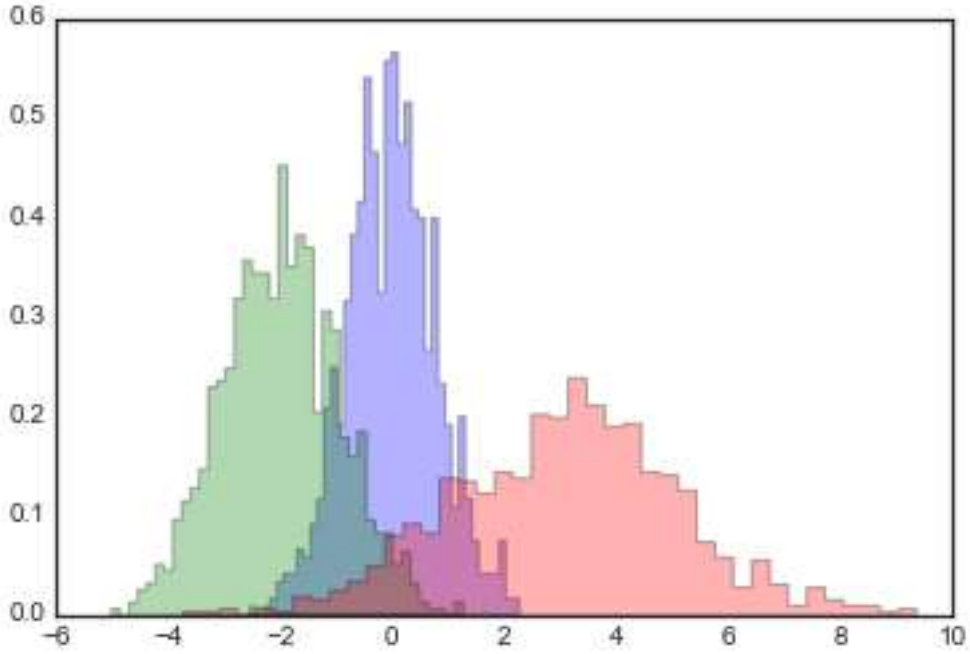


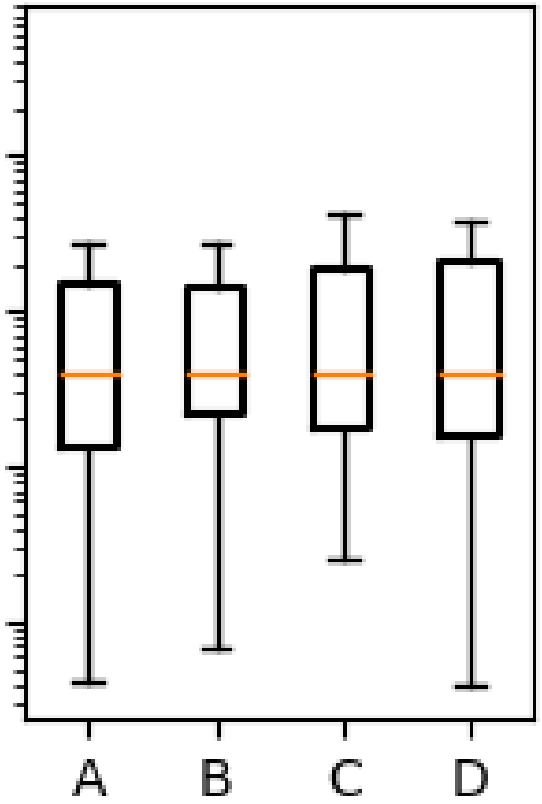






Is the difference significant?





Statistical Tests

- A wide number of statistical tests exist depending on the hypothesis and the measured data

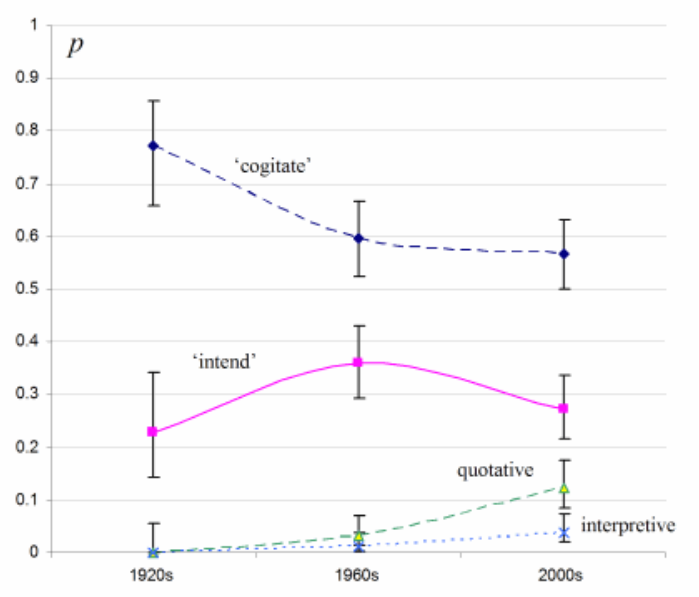
- According to the data
  - Parametric tests
  - Non-parametric tests

- According to the hipothesis
  - Comparison between two samples
  - Comparison between three samples or more
  - Just noticeable differences

Statistical Methods

➢ Parametric Tests

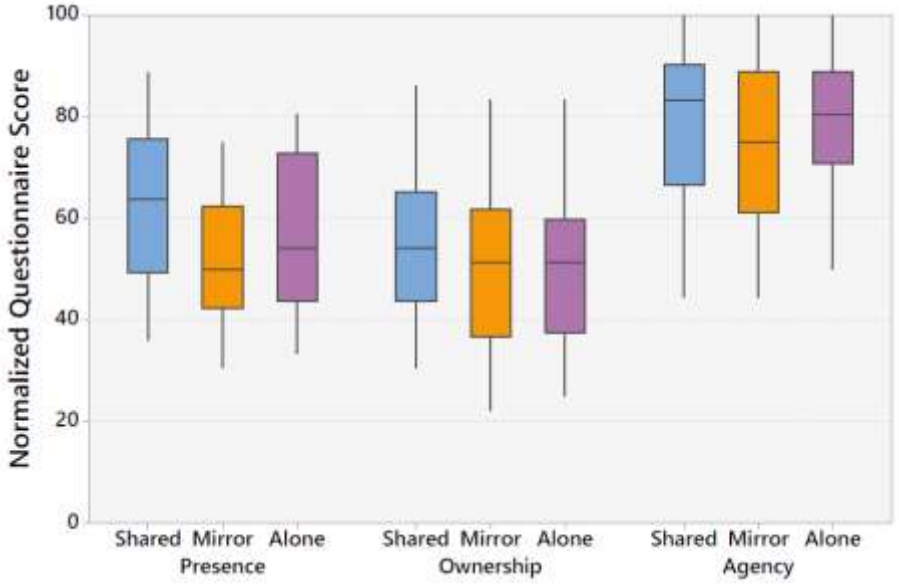
- Assume data normality – the distributions of the residuals are normal.
- Assume equality (or "homogeneity") of variances, called homoscedasticity
- More powerful (use raw data)
- Descriptive statistics: Mean
- More relevant plot: Mean plot with confidence intervals



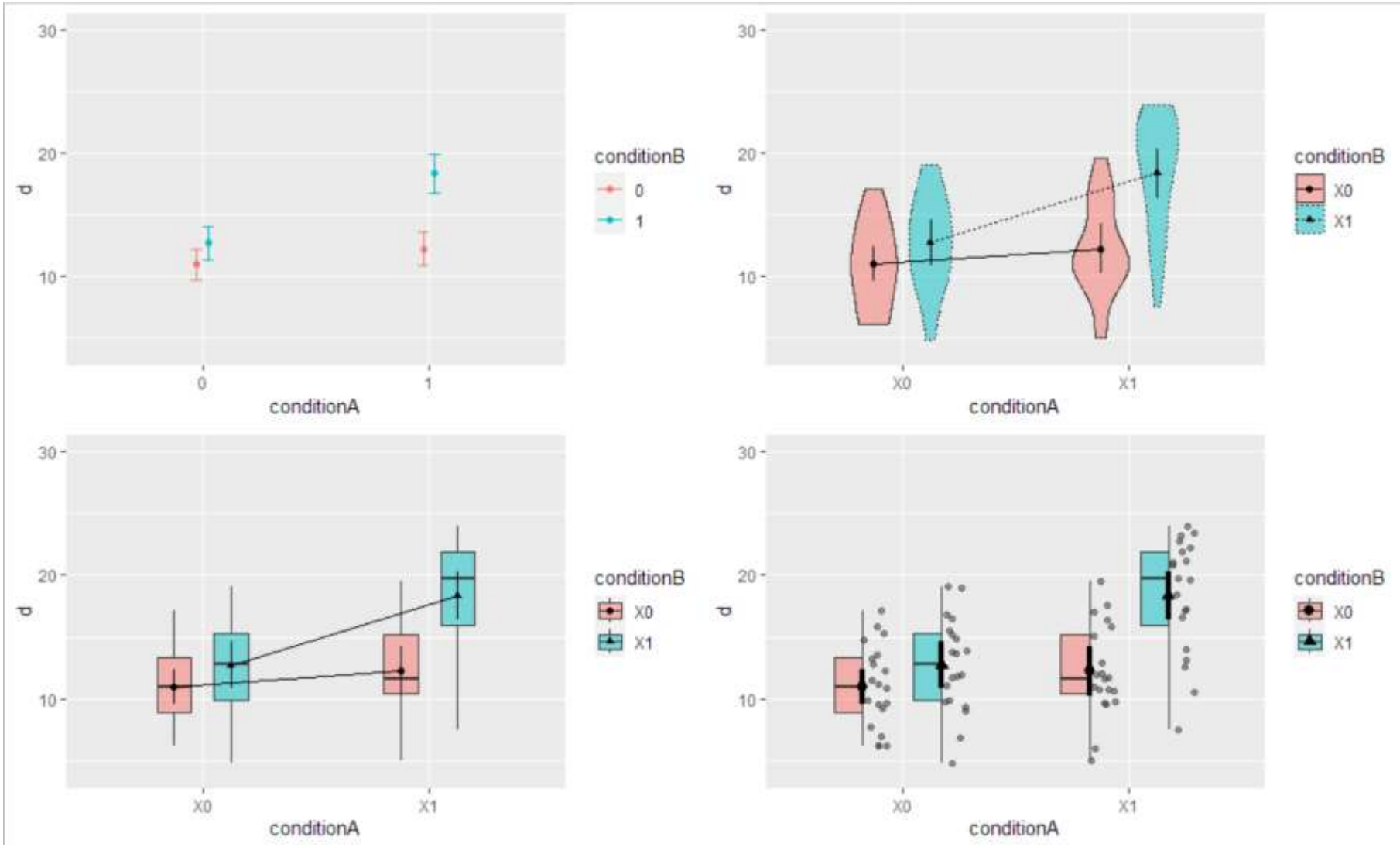
Statistical Methods

➢ Non-Parametric tests

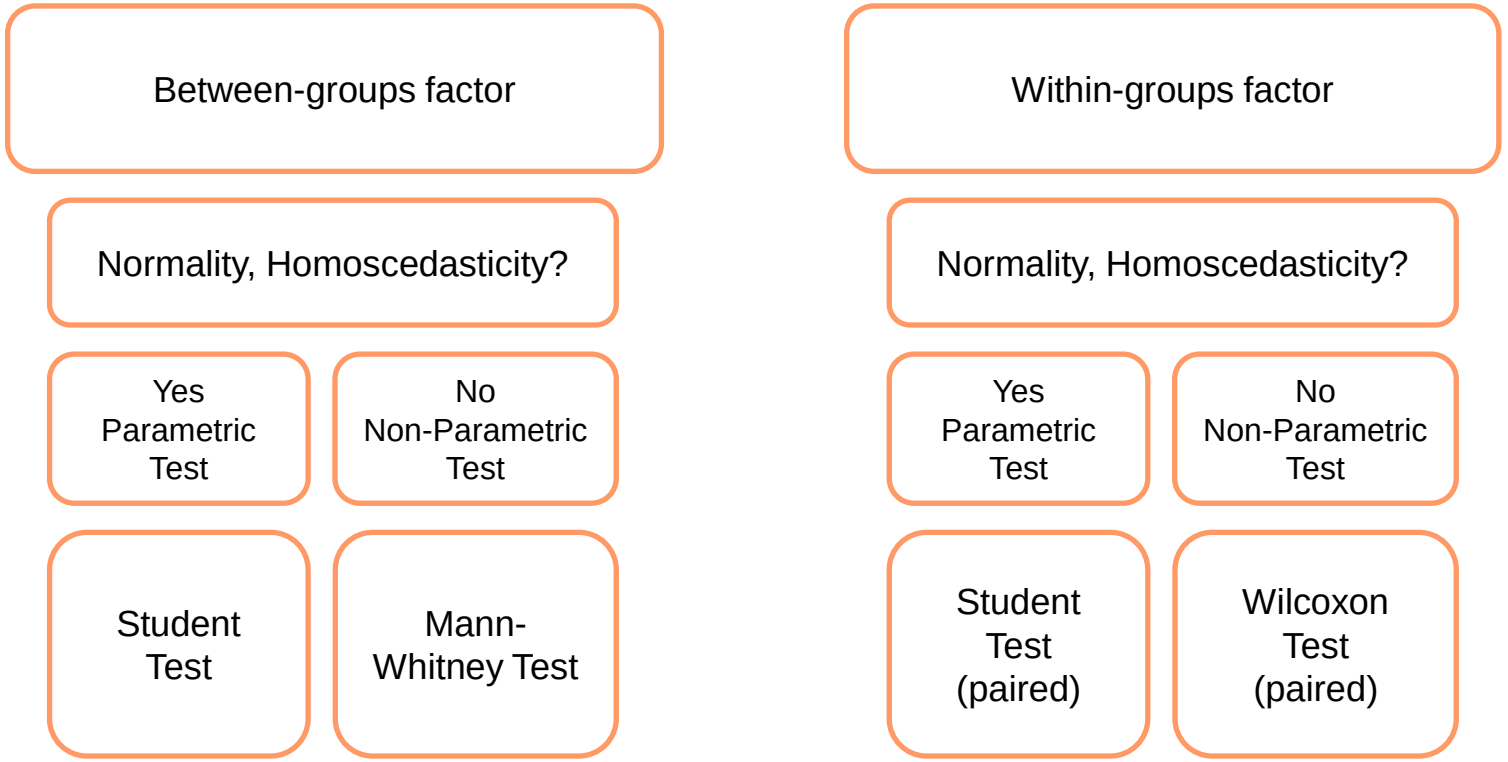
- Use in case that parametric tests are not well-suited (e.g. data from questionnaires)
- Can be used for small population sizes
- Less powerful (uses ranks or frequency of observations)
- Descriptive statistics: Median
- More relevant plot: Box plot



Other plots



One factor, two levels



One factor, three levels or more

	Between-groups factor		Within-groups factor	
	Normality, Homoscedasticity?		Normality, Sphericity?	
	Yes Parametric Test	No Non-Parametric Test	Yes Parametric Test	No Non-Parametric Test
	One-Way ANOVA	Kruskall- Wallis Test	Repeated Measures One-Way ANOVA	Friedman Test
Post hoc Tests!!!!	Tukey / Bonferroni	Mann- Whitney tests	Tukey / Bonferroni	Wilcoxon Test (paired)

Psychophysics

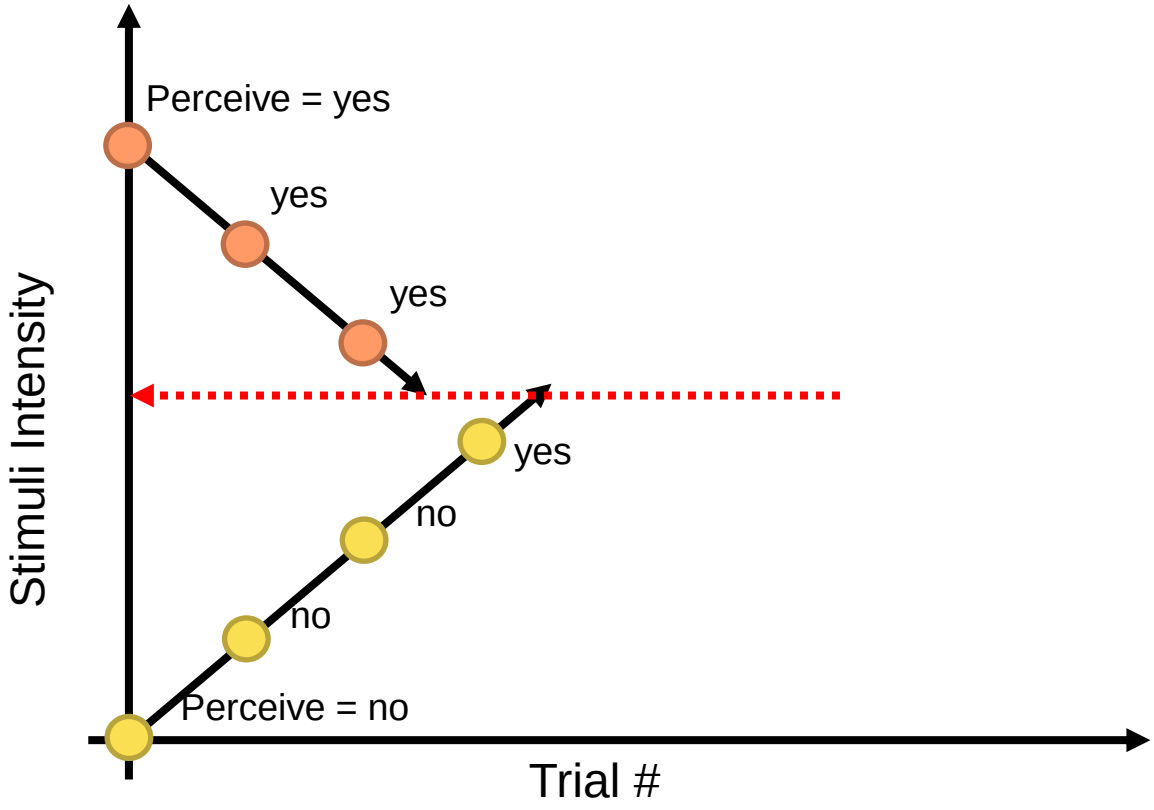
- Sometimes we would like to measure the actual difference instead of knowing if a difference exist

  - « Psychophysics quantitatively investigates the relationship between physical stimuli and the sensations and perceptions they produce.”
  - E.g. haptic perception, speed perception, color perception.

- Classical methods
  - Limit, adjustment, constant stimuli …
- Adaptative methods
  - Staircase, bayesian, effect estimation, …

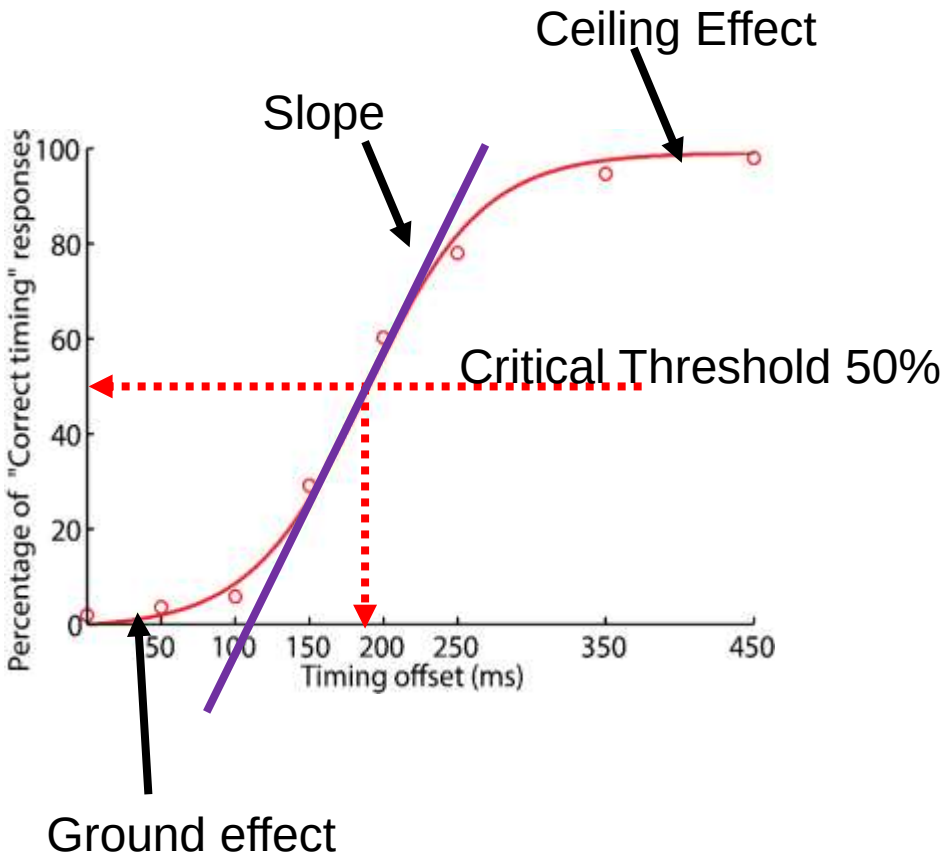
Classical Methods

- Method of the limits
  - The stimuli presented increases or decreases along the experiment until a change in the user response is measured.



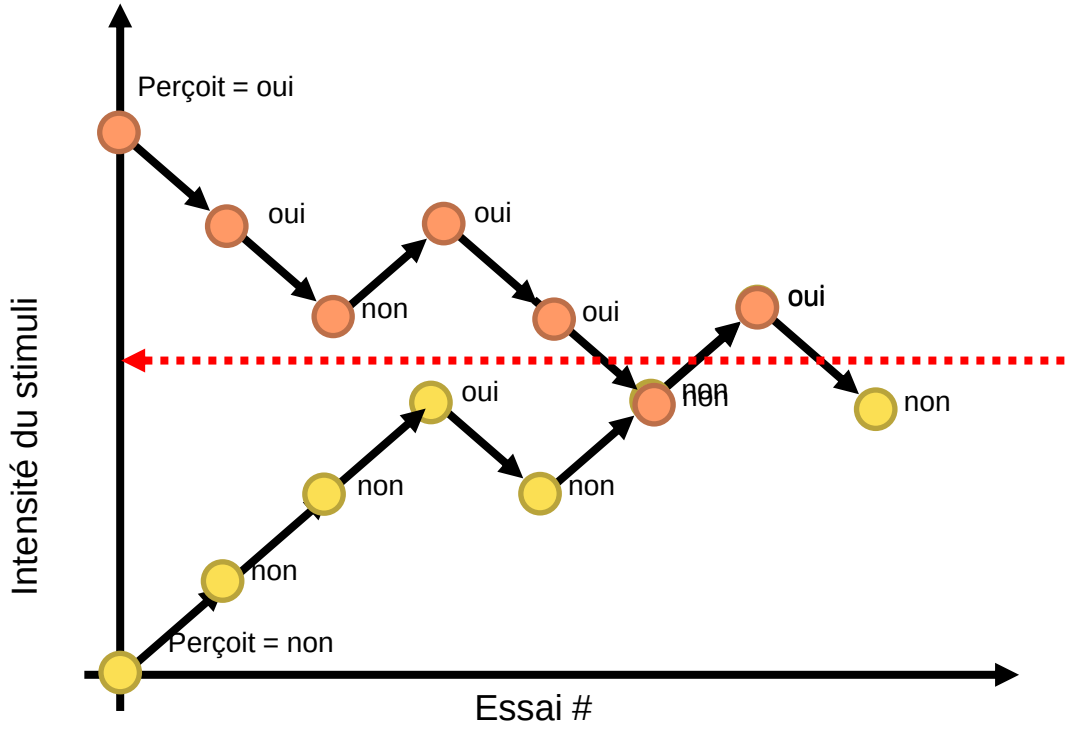
Classical Methods

- Constant Stimuli Method
  - The experimenter presents the stimuli in a random order (multiple repetitions)



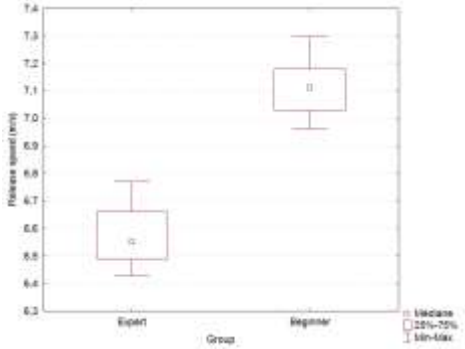
Adaptative Methods

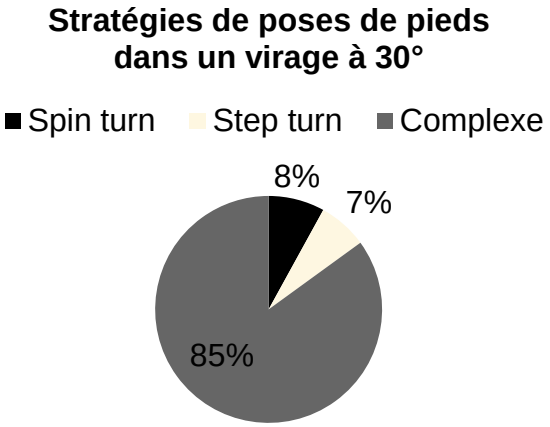
- Staircase Method
  - Similar to the limit method but each time that the participant changes its response the direction of the staircase is inversed.

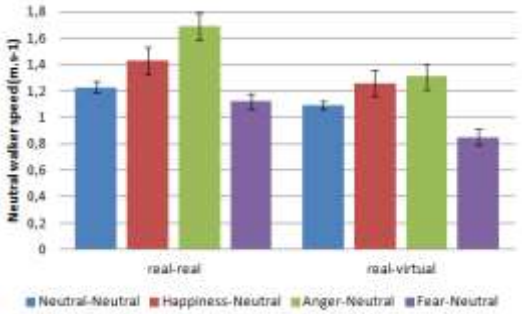


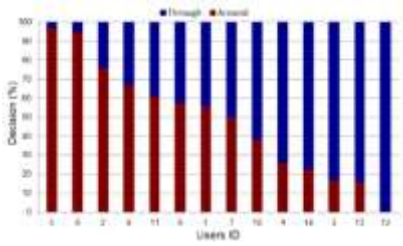
Display your results

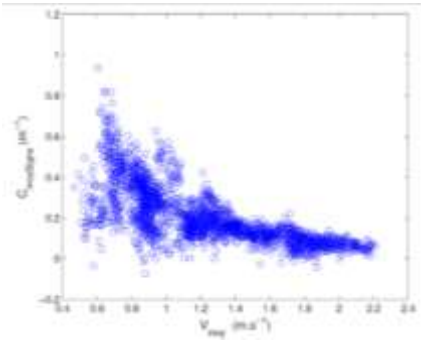
- A wide range to visualize your results
  - From the simplest …





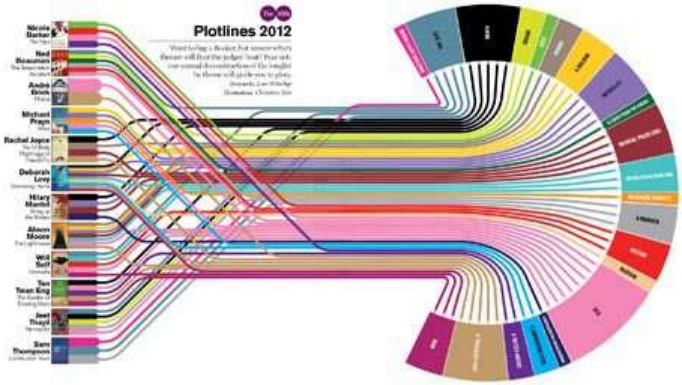




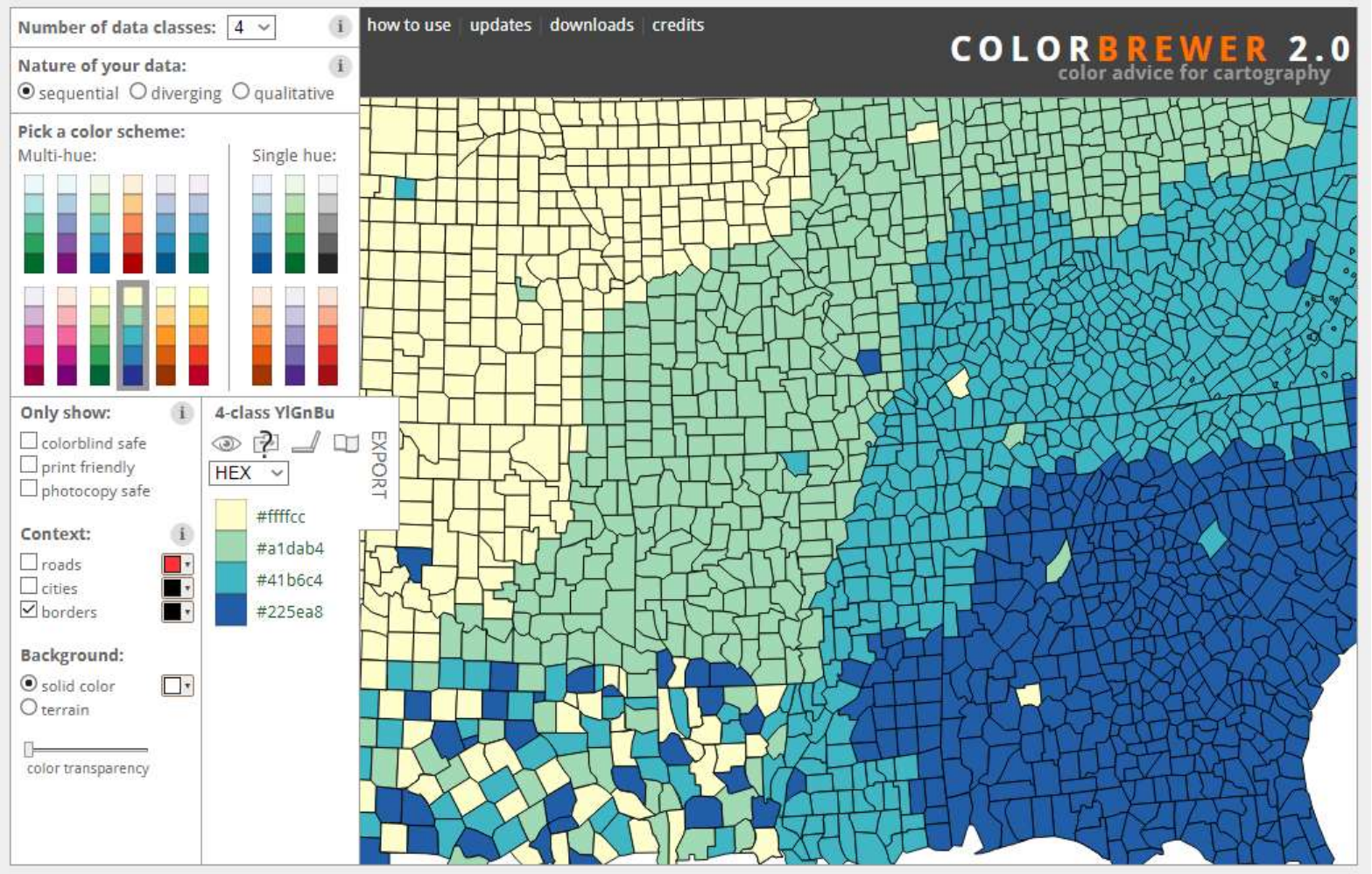


Display your results

- A wide range to visualize your results
  - … to the most complex



Color Encoding



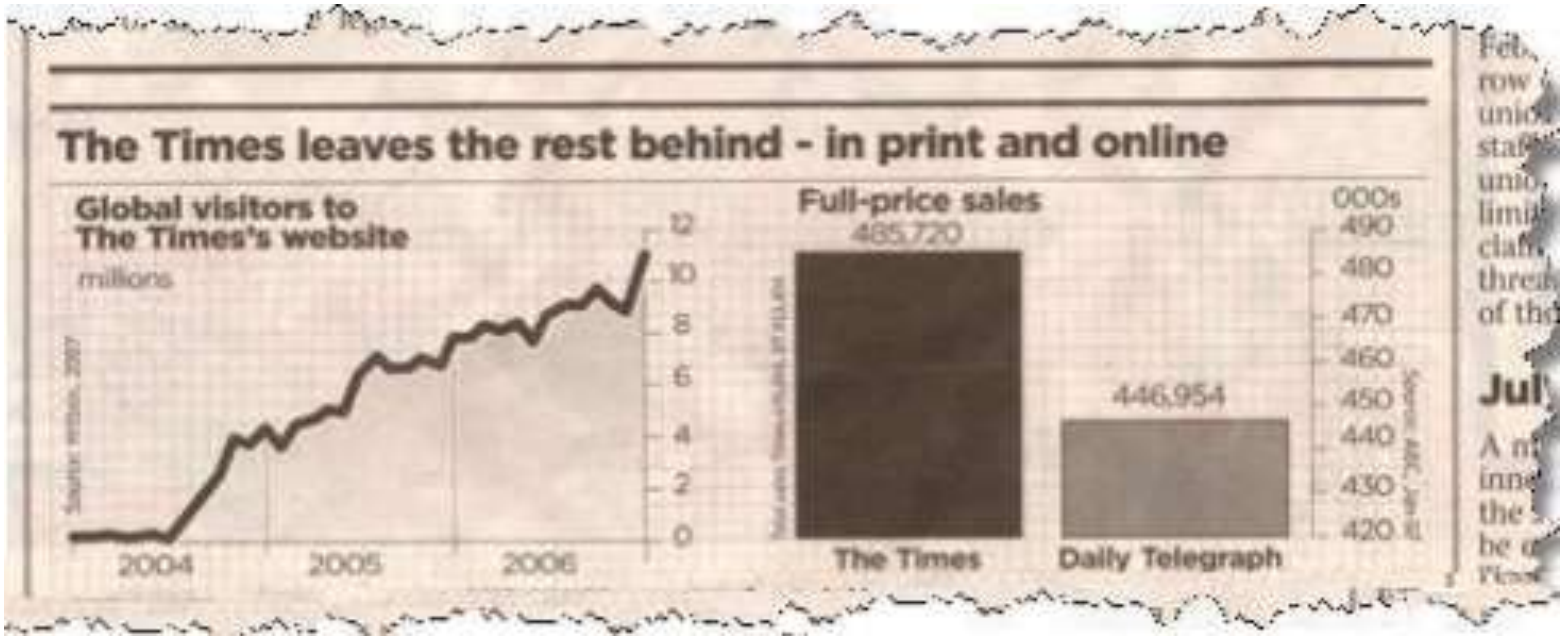
http://colorbrewer2.org/

Color Use Guidelines for Data Representation, Brewer, C. A. 1999. Proceedings of the Section on Statistical Graphics, American Statistical Association, Alexandria VA. pp. 55-60.

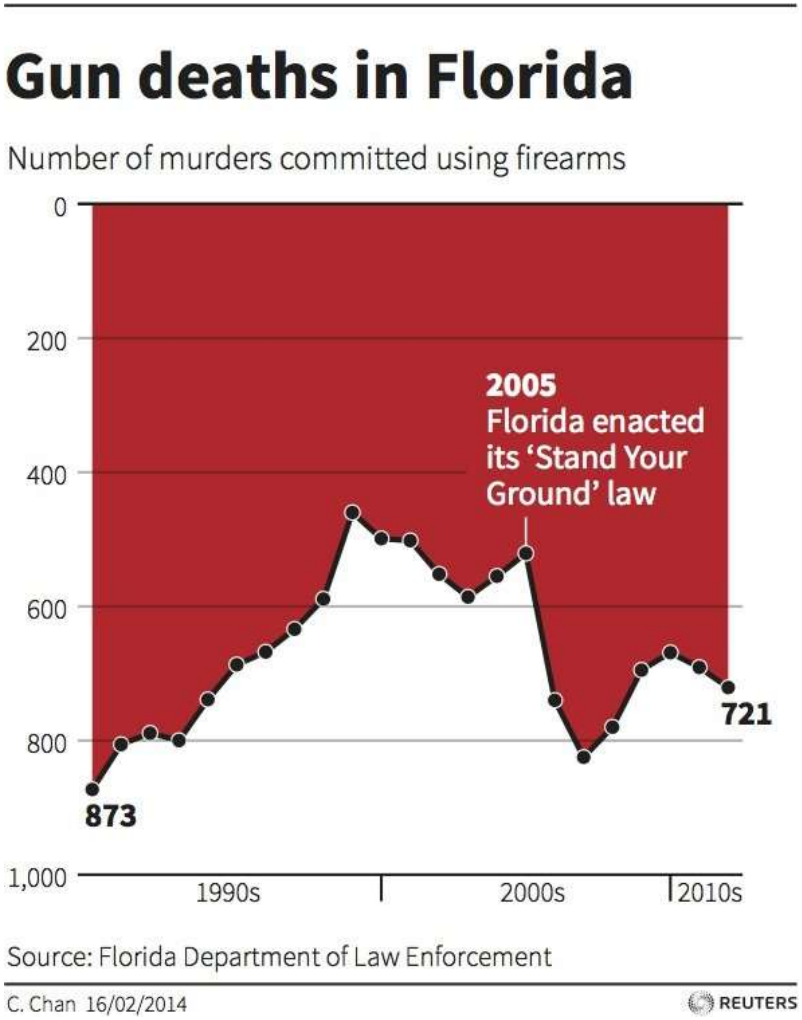
Avoid misleading plots

- Avoid non useful plots

Avoid misleading plots



Avoid misleading plots



Statistical Ressources

- Software

SigmaSTAT
Advisory Statistics for Non-statisticians

R

StatSoft
STATISTICA

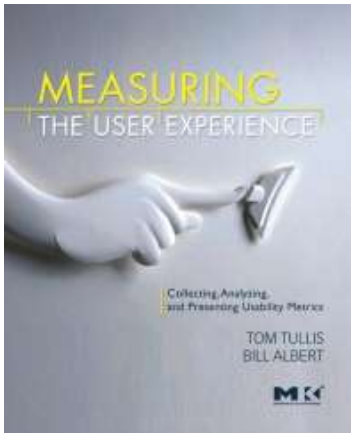
SP

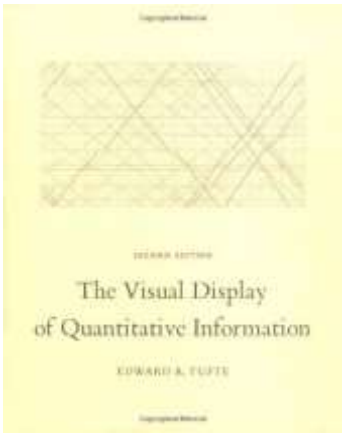
psignifit 3.0β

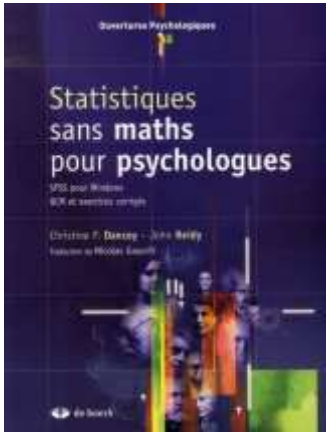
XLSTAT

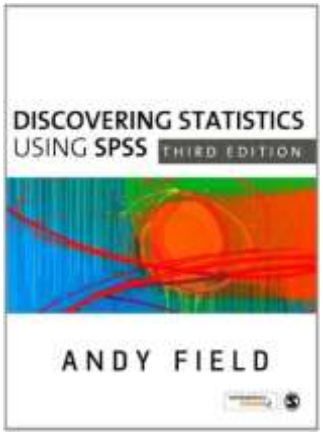
MATLAB
The Language of Technical Computing

- Further reading









www.statisticshell.com

User Interface Evaluation

- Introduction
- Evaluation tools
- Evaluation methods for 3D UI
- Evaluation metrics
- Evaluation methodology
- Challenges for 3DUI/VR evaluations

Challenges for 3DUI/VR evaluations

- Physical Environment Issues
  - Use of non-traditional I/O devices
  - Users may be standing rather than sitting
  - Variable interaction space

- Examples
  - HMD/CAVE : users can bump into walls, trip over cables
  - Most 3D displays do not support simultaneous viewers
  - Video recording of both the user and the interface
  - Collaborative systems

Challenges for 3DUI/VR evaluations

- Evaluator Issues
  - A user study might require several evaluators
    - 3DUI hardware and software are less robust
    - Need of simultaneously process multimodal input
    - Need of different competences
- User Issues
  - Potentially strong variability → Need to increase sample size
  - Hard to differentiate novice from expert users
- Lack of verified guidelines

Challenges for 3DUI/VR evaluations

- Simulator **Sickness** (especially for VR)
  - No exposure should last more than 20 minutes continuously
  - If experiment longer than 20 minutes, plan rest breaks
  - Ask subject often how they are feeling
  - Allow subjects to quit anytime they want
  - Measure levels of discomfort several times during long experiments
  - Warn subjects not to drive immediately afterwards if they experience strong symptoms

Guidelines for 3DUI/VR Evaluation

- Begin with **informal** evaluations
  - Experts and novices
  - Identify big flaws of the system
- Perform **pilot** studies to ensure the viability of the study
  - User studies are long and potentially expensive
- Consider multiple evaluation **metrics**
  - Objective and subjective
  - Gather as much information as possible. Data is your precious!!
- Consider **interactions** between factors
  - A single technique will not be the best for all situations

Wrap-up

- Introduction
- The User in the Loop
- Interacting with virtual worlds
- Evaluation of user interfaces
  - Concepts and definitions
  - Evaluation tools
  - Evaluation methods
  - Evaluation metrics
  - Evaluation methodology
  - Challenges for 3DUI/VR evaluations

The End