

Evaluation in Visualization

Michael Sedlmair

Adapted by Eliane Zambon Victorelli

Outline

- Motivation
- Forms of evaluation
- Methods for evaluating technique-driven projects
 - Algorithmic Performance & Image Quality
 - Controlled Experiments
- Methods for evaluating problem-driven projects
 - Observations/Interviews
 - Usability Testing & Prototyping
 - Case/Field Studies
 - Adoption rates

Motivation

What is Evaluation?

What is Evaluation?

“Evaluation is the systematic assessment of the worth or merit of some object”

(Quasi-standard definition from 50s/60s)

Why Evaluating?

- To ensure quality in product development
- To compare solutions
- To provide quantitative results
- To get a scientific statement (instead of personal opinion)
- To convince your audience

Evaluation in Vis is related to ...

- Computational Performance
- HCI (e.g., Usability)
- Perceptual Psychology
- Cognitive Reasoning/Sense-making
- Social Science

- ...and of course: Statistics!

Evaluation comes in many forms...

- Who: With Users vs. Without Users?
- Why: Formative vs. Summative?
- How: Quantitative vs. Qualitative?
- Where: Field studies vs. Lab studies?
- When: Before vs. During vs. After Development

Who: With Users vs. Without Users?

Evaluation through expert analysis:

- the designer or HCI/VIS expert
- identify violation of principles
- any stage in the development process
- relatively cheap

Evaluation through user participation

- the people for whom the system is intended
- it can be expensive
- at least a prototype.

Why: Formative vs. Summative?

Formative:

- goal: inform the design
- carried out throughout the design process
- prototypes at different fidelity

Summative:

- goal: product has the desired levels of quality
- carried out at the end of a design process
- system implemented or high-fidelity prototype

Why: Vis Study Goals

Understand perceptual and cognitive principles	Understand human perceptual or cognitive characteristics, often by measuring performance at abstract tasks
Understand context	Understand the context in which a visualization will be used, user characteristics, tasks, environment, social context, work practices, communication.
Compare tools or visualization and interaction techniques	Compare two or more approaches to identify strengths and weaknesses of each or to validate a hypothesized improvement over a baseline design
Evaluate one tool or technique	Identify strengths, weaknesses, and/or limitations of one single approach

Tory, M. (2013). User studies in visualization: A reflection on methods. In Handbook of Human Centric Visualization (pp. 411-426).

New York, NY: Springer New York.

How: Quantitative vs. Qualitative?

	Quantitative	Qualitative
Objective	The chip speed of my computer is 20GHz	"Yes, I own a computer"
Subjective	On a scale of 1-10 my computer scores 7 in terms of ease of use	I think computers are too expensive.

	Quantitative	Qualitative
Objective	Speed, accuracy	Factual observations of user actions
Subjective	Likert scales	Free responses

Likert scales

A. The system was easy to use.

1 – Strongly disagree 2 – Disagree 3 – Neutral 4 – Agree 5 – Strongly agree

B. The system was fun to use.

1 – Strongly disagree 2 – Disagree 3 – Neutral 4 – Agree 5 – Strongly agree

C. The system was easy to learn.

1 – Strongly disagree 2 – Disagree 3 – Neutral 4 – Agree 5 – Strongly agree

How: Empirical Approach

Quantitative experiment	Makes a direct comparison between two or more controlled conditions and measures a quantitative difference between them.
Qualitative observ. study	Answers exploratory questions using mainly qualitative data gathered through techniques such as interview or video.
Inspection	A small number of experts inspect visualization tools or techniques using a pre-defined protocol.
Usability study	Users complete tasks with a visualization tool, technique or interaction method to assess whether it meets specified criteria.

Various Evaluation Methods used in Visualization

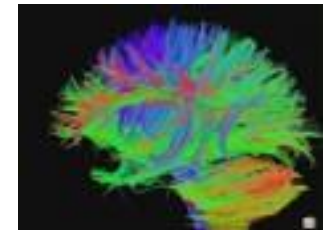
- Usability Testing
- Rapid Design Feedback / Prototyping
- Controlled User Experiments
- Field Studies
- Algorithmic Performance & Image Quality
- ...
- **When to use which?**

Two flavors of visualization projects

- technique-driven
- problem-driven

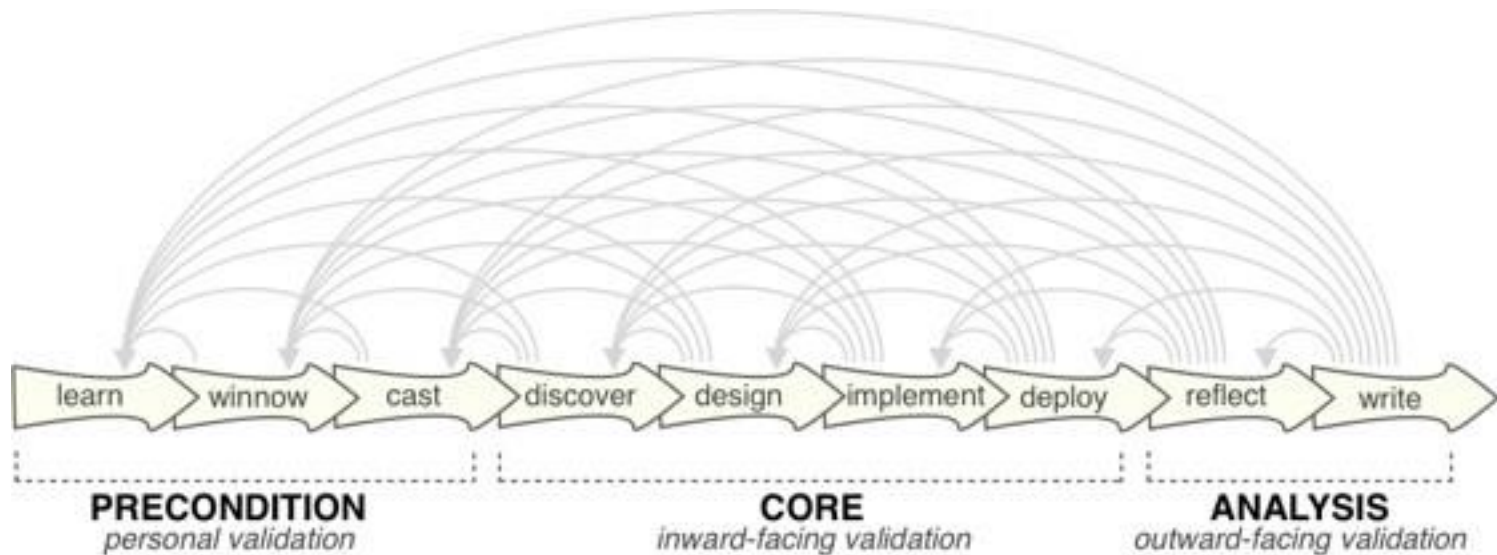
Technique-driven

- Scenario: “The new **thing** we built is **good/better** than what we currently have”
- Thing:
 - visual encoding/interaction technique
 - algorithm
- Good/better:
 - time (runtime, user performance)
 - errors
 - subjective user preference
- Evaluation usually at the end of the Project
- Example



Problem-driven

- Scenario: Working together with users that have data and problems, designing a VIS tool for them, and validating the design.



Sedlmair et al. (InfoVis 2012): **Design Study Methodology:**
Reflections from the Trenches and the Stacks

Problem-driven: Evaluation along the entire design process

Do I understand the problem?

- Who are my target users?
- What is their problem?

Designing a tool

- Addressing users' needs?
- Usable?

Is the solution good/better?

- See technique-driven metrics
- How is the system used in the wild?
- Are people using it?
- Does the system fit in with existing workflow?

Methods

(often used in technique-driven projects)

Algorithm Performance & Image Quality

Algorithmic Performance & Image Quality

- Complexity
 - measured in terms of size of input problem
 - e.g. input size of a volume is not N , but N^3
- Scalability
 - closely related to Complexity
 - Investigate interactivity (speed)
 - Investigate resource constraints (memory)
- Image Quality

Image Quality

- Quantitative: Metrics
- E.g., metrics for graph drawing
 - line crossings?
 - area?
 - sum of edge length?
 - uniform edge length?
 - ...










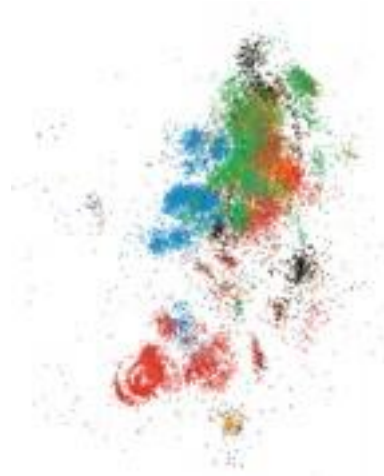
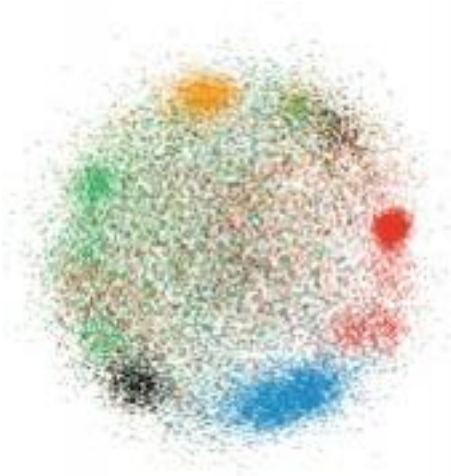
	R_1	R_2	R_3	R_{4a}	R_{4b}	R_{5a}	R_{5b}
N1 GEM 	1	1	0.68	0.75	0.57	0	0.62
N2 EX TDAG 	0.98	0.78	0.52	0.65	0.65	0	0
N3 TS 	1	0.83	0.64	0.75	1	0	0
N3 DAG 	1	0.83	0.75	0.63	0.33	0.67	0
N4 TS 	0.96	0.82	0.33	0.64	0.98	0	0.18
N4 TSC 	0.94	1	0.02	0.43	0.23	0	0.11
N5 TS 	0.97	0.81	0.57	0.64	1	0	0.29
N5 GEM 	0.95	1	0.01	0.53	0.48	0	0.44
N6 TS 	0.96	0.82	0.33	0.64	1	0	0.29
N6 KAM 	0.93	1	0.01	0.49	0.52	0	0.42

Figure 12. Examples of the application of the aesthetic metrics: North graphs with GraphLet algorithms applied.

Image Quality

- Qualitative Discussion of results of an algorithm
- *Bad*: Look at the great image!
- *Good*: The image clearly shows five separable clusters with different colors, while the others ...



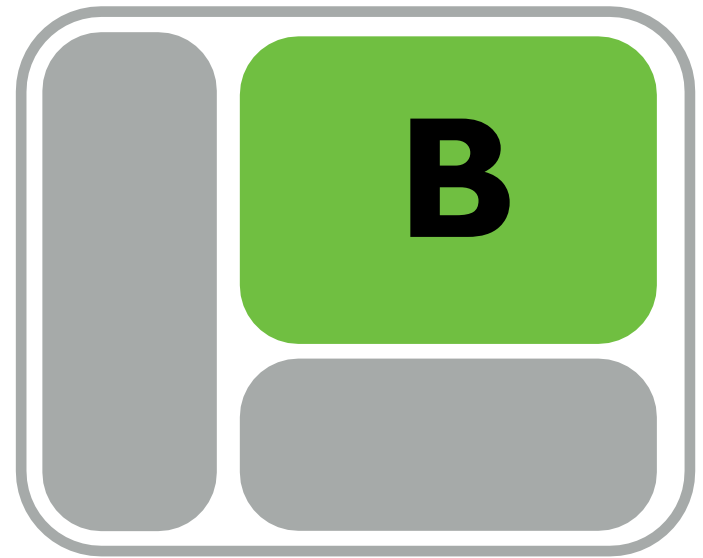
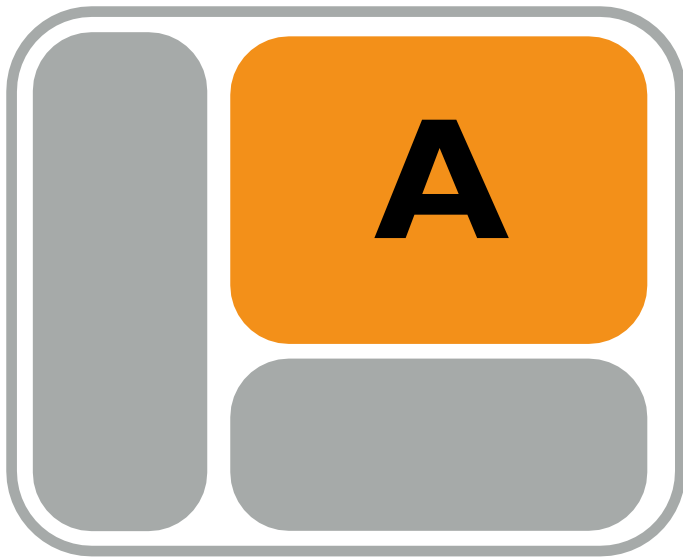
User Performance & User Experience

Controlled experiment

- Other names:
 - Laboratory Experiment
 - Lab study
 - User Study
 - A/B Testing (used in marketing)
- Most common goal:
 - Is your novel technique better than state-of-the-art techniques? (faster? / less errors? / more insights? ...)

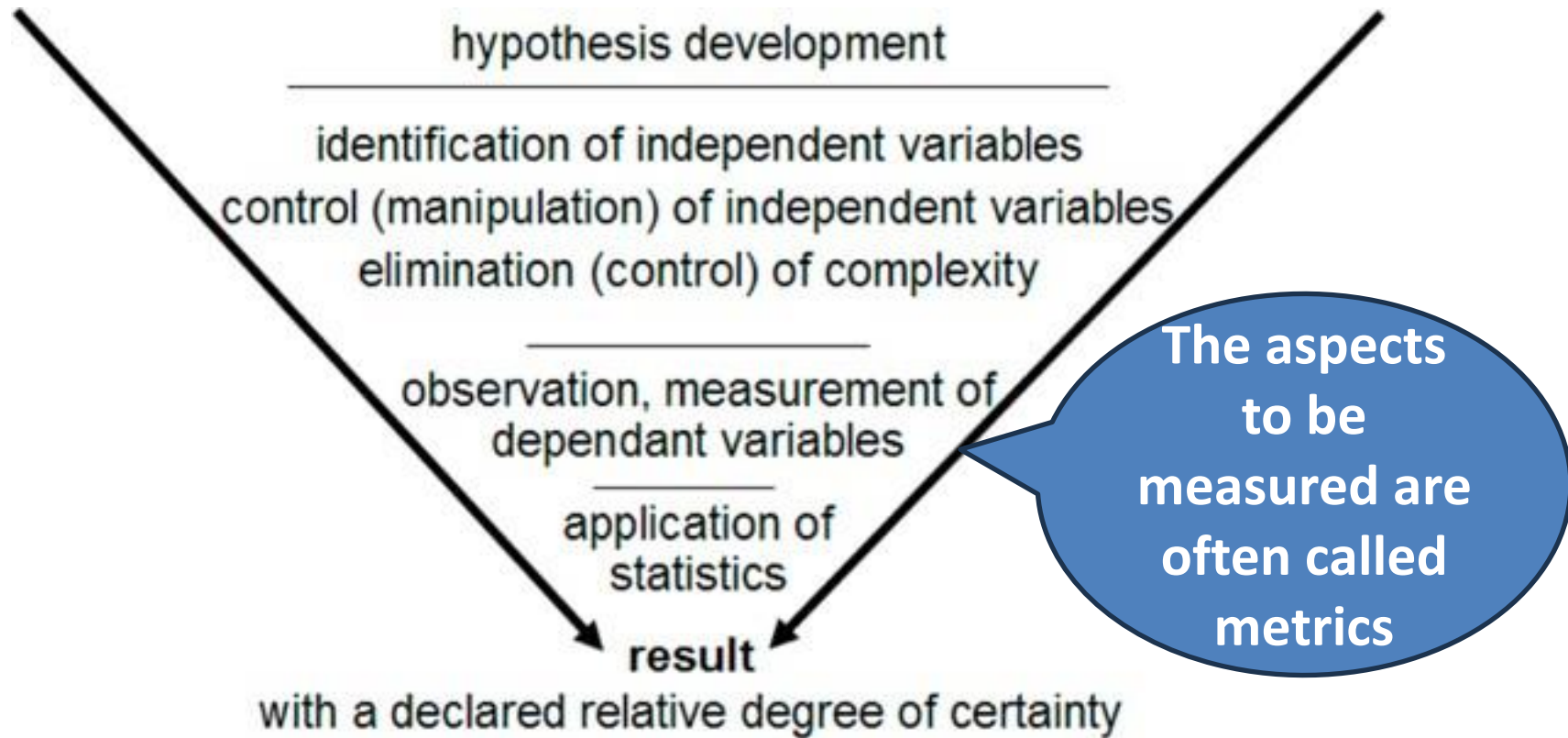
Scenario

Your cool new vis



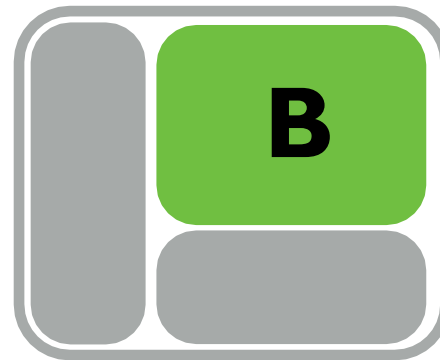
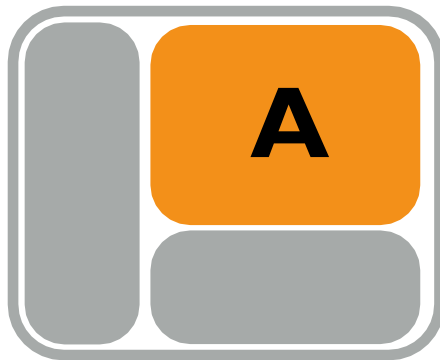
Which is better?

Test it with users!



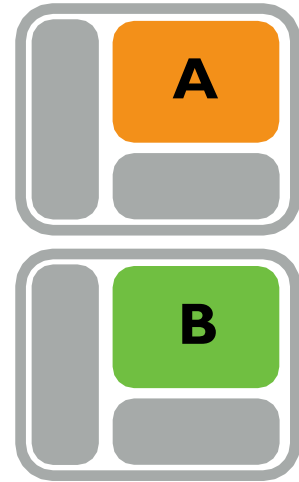
Step 1: Develop Hypothesis

- A precise problem statement
- Example:
 - $H_1 =$ For a given task, participants will be faster when using visualization A than visualization B
 - Null-Hypothesis $H_0 =$ no difference in speed



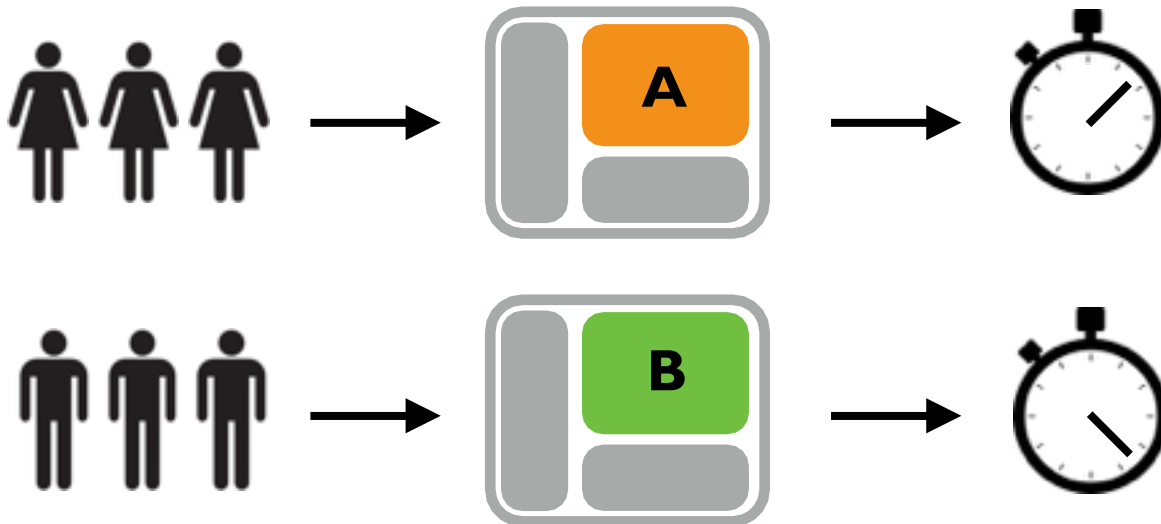
2: Independent Variables

- Factors to be studied
- Typical independent variables (in Vis)
 - Different visualization design
 - Different interaction techniques
 - Task type: e.g., searching/browsing
 - Participant demographics: e.g., experts/non-experts
- Control of Independent Variable
 - Levels: The number of variables in each factor
 - Limited by the length of the study and the number of participants
- How different?
 - Entire interfaces vs. very specific parts



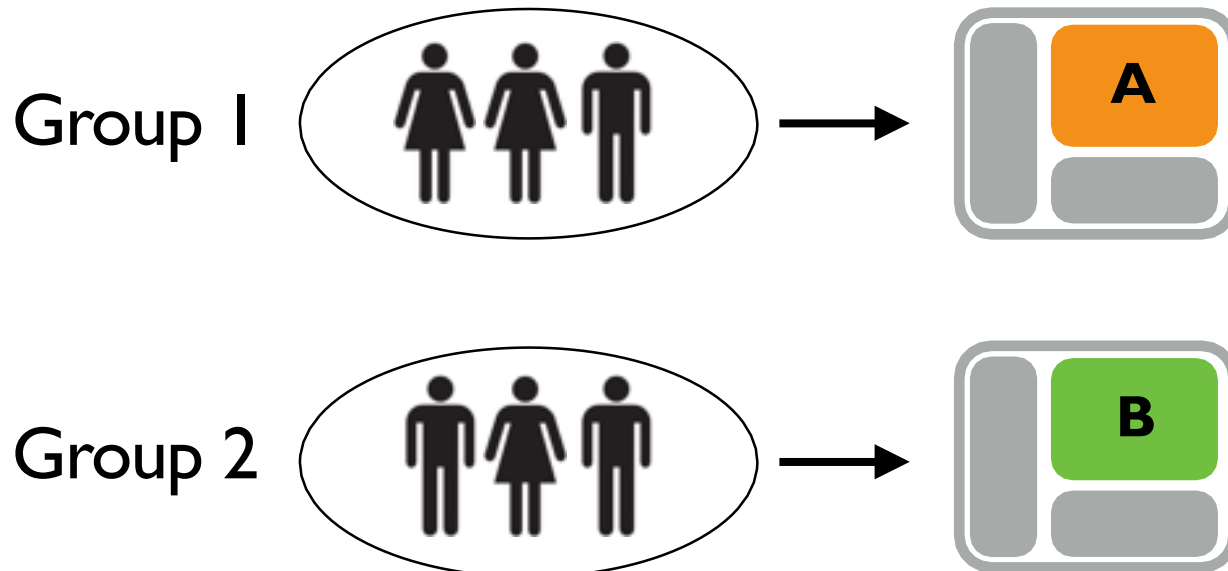
3. Control Environment

- Make sure nothing else could cause your effect
- Control confounding variables
- Randomization!



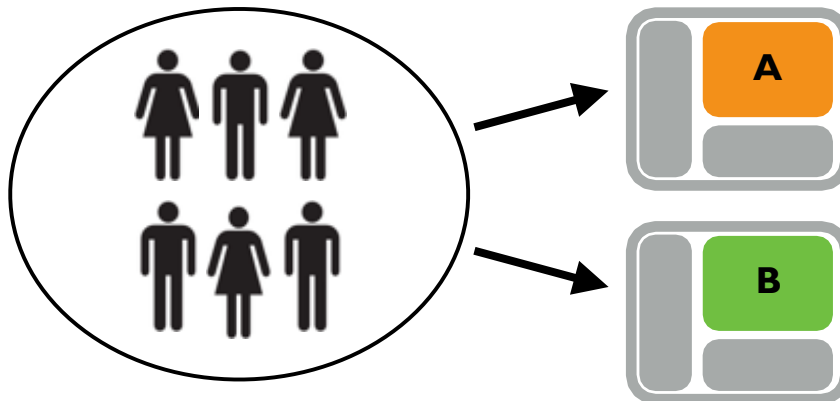
Different Designs: Between-Subjects

- Divide the participants into group, and each group does some of the conditions
- Randomize: Group Assignment
- Potential problem?



Different Designs: **Within-Subjects**

- Everybody does all the conditions
- Can account for individual differences and reduce noise (that's why it may be more powerful and requires less participants)
- Can lead to ordering effects —> Randomize Order



4. Dependent Variable

- What is “better”
- Metric - The things that you measure
- Performance indicators:
 - task completion time, error rates
 - accuracy
- Subjective participant feedback:
 - satisfaction ratings,
 - Questionnaires, interviews
- Observations:
 - behaviors, signs of frustrations...

Results: Application of Statistics

- Descriptive Statistics

- Describes the data you gathered (e.g. visually)

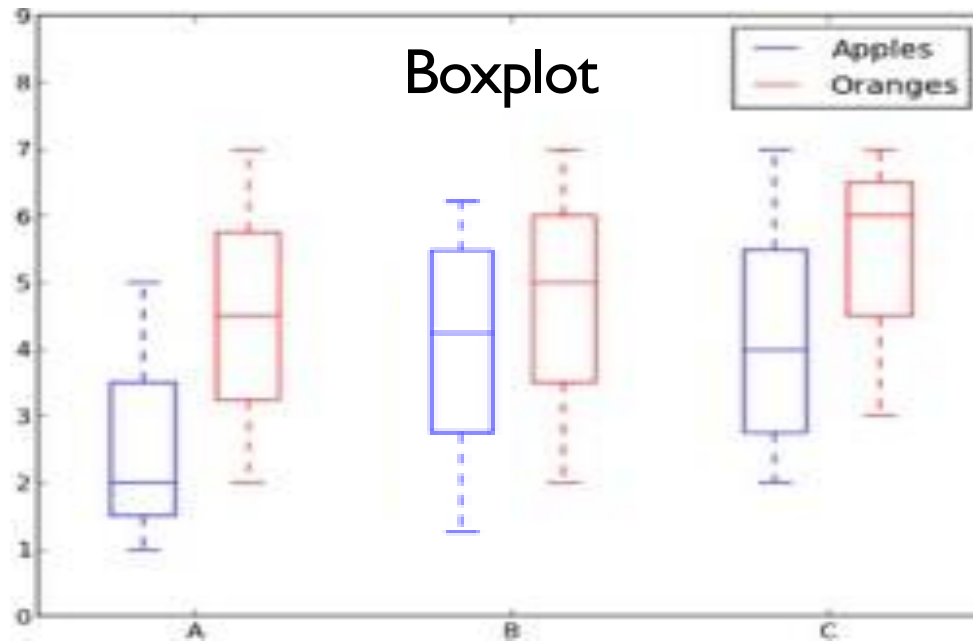


- Inferential Statistics

- Make predictions/inferences from your study to the larger population



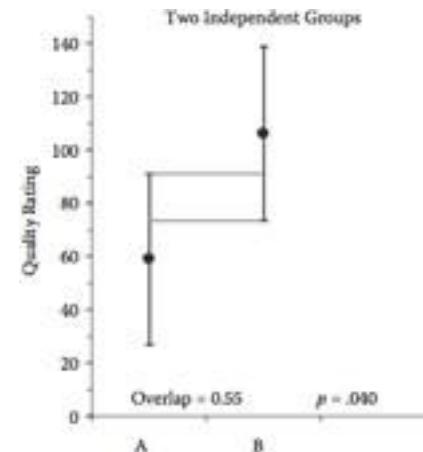
Descriptive statistics



- Mean
- 25/75% Quartiles
- Min / Max
- (alternative: with outliers)

Inferential statistics

- Goal: Generalize findings to the larger population
- Statistically significant results
 - $p < .05$
 - The probability that we incorrectly reject the Null-Hypotheses
- Many different tests
 - t-test, ANOVA, ...



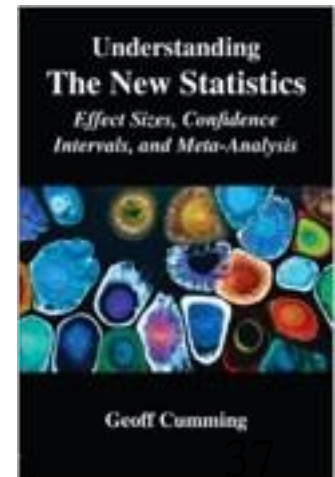
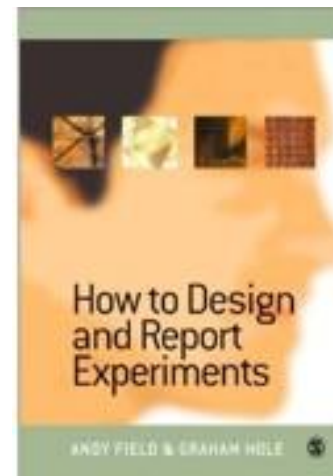
Pragmatic concerns

- Number of participants
 - Depends on effect size and study design — power of experiment
 - Usually 15-20 (per group)
- Recruiting participants
 - Reflecting larger population?
 - Lab study vs. mechanical Turk



Pragmatic concerns

- Possible confounds?
 - *Learning effect*: Did everybody use the interface in a certain order?
 - *Biasing*: "my visualization vs. this other visualization",
- Don't compare to 'dead horses'!
- Pilot studies
 - Should test the study setup for possible problems
- Ethics

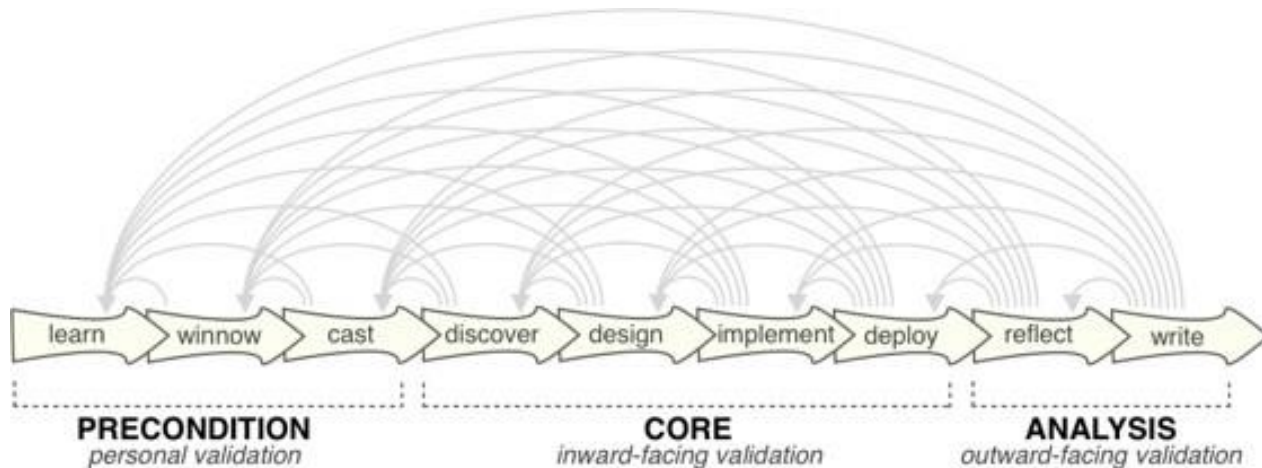


Methods

(often used in problem-driven projects)

Important Milestones

- Problem Characterization:
 - Understand Task and Problem Understanding
- Design:
 - Iterative Process and Usability Engineering
- Validation of final visualization tool



Problem Characterization

Goals

- What kind of problems is the system aiming to address?
- Who is your target users?
- What are the tasks? What are the goals?
- What are their current practices? What tools do they use?
- Why are these tools not solving the problem at hand?
- Why and how can visualization be useful?
(e.g., presentation, communication, debugging, speeding up the workflow, hypothesis testing/creation,...)
- **Evaluate with users!**

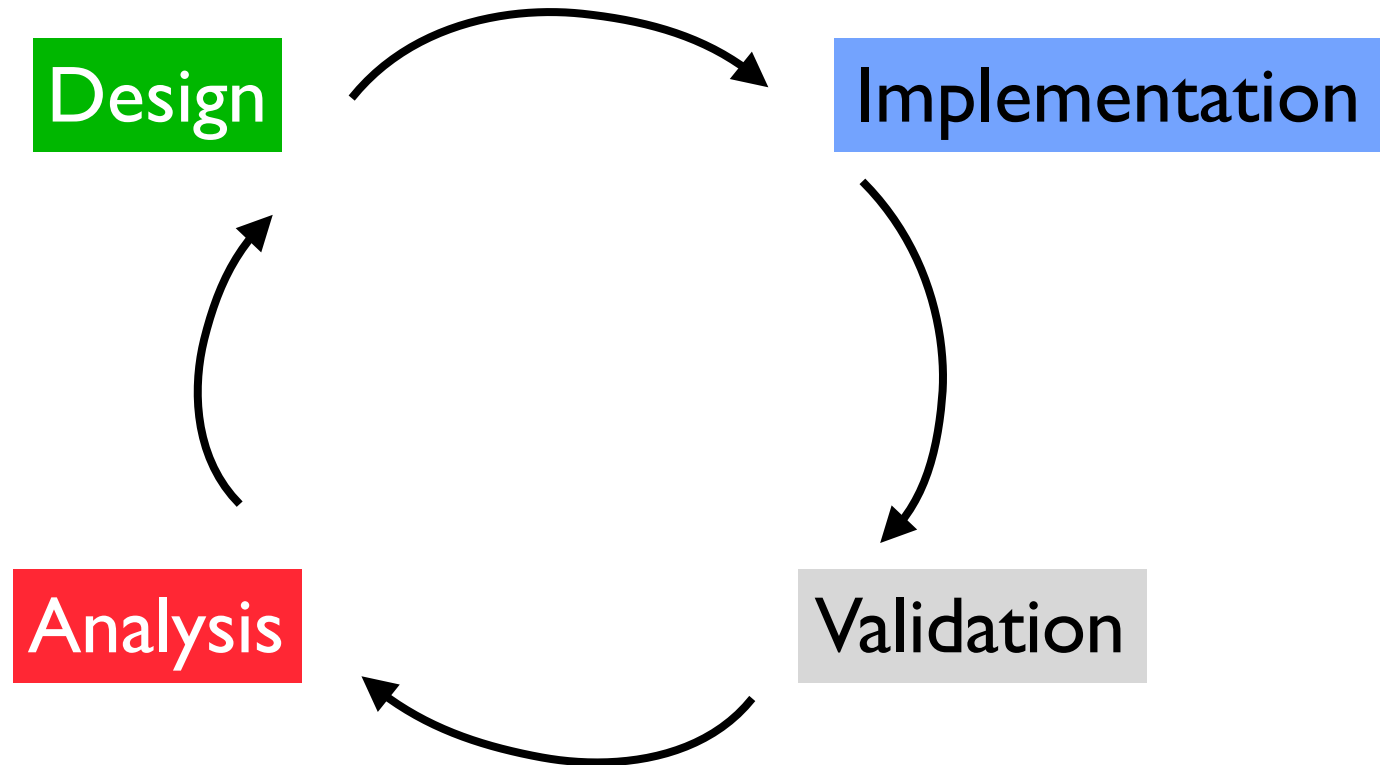
Qualitative Methods

- Observation Techniques
 - In Situ Observations (fly-on-the-wall)
 - Participatory Observations
 - Laboratory Observational Studies
- Interview Techniques
 - Contextual Interviews
 - Focus Groups

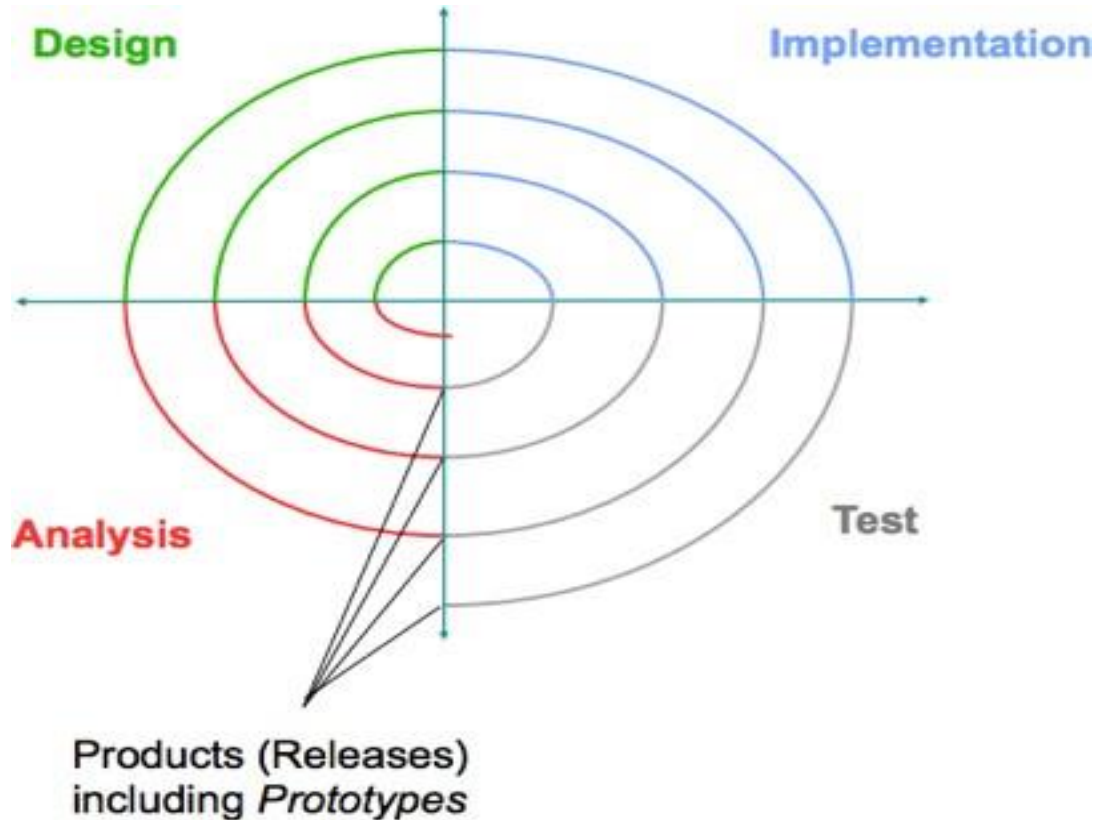


Iterative Design

2. Iterative Design



Rapid Prototyping



Increasing fidelity of prototypes

Paper Prototypes & Data Sketches

- Interviews, Observations, ... with Prototypes
- “Above all, show them their data”



Lloyd and Dykes (InfoVis 2011): “Human-centered approaches in geovisualization design: Investigating multiple methods through a long-term case study.”

Evaluation Goals

- Is the tool usable?
- Improve product design!
- 5 E's: Is the interface ...
 - Effective?
 - Efficient?
 - Engaging?
 - Easy to learn?
 - Equally usable by different groups?

Methods: Usability Testing

- Observation
 - enables observers to see first-hand the process of task completion
 - drawback: can't see what users think
- Think aloud protocol
 - participants think aloud as they are performing a set of specified tasks
- Note taking, audio-, and/or video-recording

Methods: Usability Inspection

- Without users!
- Done by Vis/HCI/domain experts
- Heuristic evaluation
 - Experts reviews an interface design with respect to a set of predefined heuristics
- Cognitive walkthrough:
 - experts ‘walk’ through an interface following a specified set of tasks.
 - Step through the task. At each step, ask yourself four questions.

Validation of Final Tool

Validation of Final Visualization Tool

- Recap — Technique-driven evaluation:
 - “The new thing we built is good/better than what we currently have”
 - algorithm speed, image quality, user performance/experience
- Potential shortcomings?

Validation of Final Visualization Tool

- Recap — Technique-driven evaluation:
 - “The new thing we built is good/better than what we currently have”
 - algorithm speed, image quality, user performance
- Potential shortcomings?
 - missing holistic view, context of use, ...
 - are people really using it

Case Studies / Field Studies

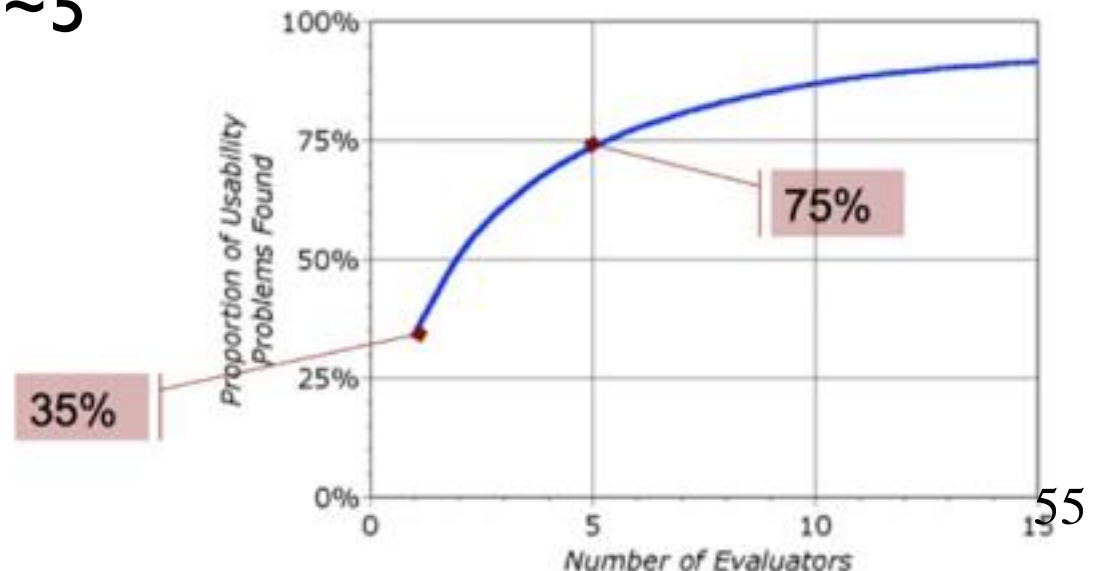
- Focus on realism
- e.g., MILC Studies = Multi-dimensional in-depth long-term Case Studies (Shneiderman 2006)
- Goal
 - reveal a richer understanding by using a more holistic approach
 - anecdotal evidence: did it really help to solve real problems?
 - examples: so far unknown insights into the data, speed-up of workflows, helped debugging,...

Adoption rates

- Do your users keep on using your tool after you stopped poking them?
- Example:
 - Checking back 3 month after a project ended
 - How many people are still using the tool?
 - How often? (Multiple times/day, once/month, ...)

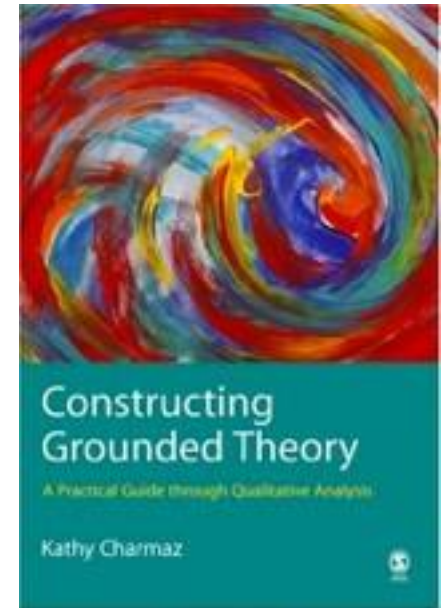
Pragmatic concerns

- Number of participants
 - Usually way less than in controlled experiments: different goals!!
 - problem/case/field studies: usually ~5 is enough, even one can be enough
 - usability tests: ~5



Pragmatic concerns

- Analyzing qualitative data?
 - Grounded theory
 - Open and axial coding
- How much rigor is necessary?
 - Rapid design iterations VS.
 - Reliable scientific findings



Summary

- Methods for evaluating technique-driven projects
 - Algorithmic Performance & Image Quality
 - Controlled Experiments
 - Methods for problem-driven projects
 - Observations/Interviews
 - Usability Testing & Prototyping
 - Case/Field Studies
 - Adoption rates
- No clear-cut line!
- Many different ways of organizing evaluation methods

Exercise: Information Visualization Evaluation Planning

You should plan a user experience assessment with the information visualizations you built in previous modules. To do this, you must carefully answer the questions in the questionnaire.

[Planning guide for user experience evaluation with information visualization](#): form with questions that help you reason about the different aspects of a visualization assessment.

[Catalog of user experience evaluation with information visualization](#): supporting material with definitions, descriptions, most common uses and examples of visualization evaluations that have already been published.

Literature

- Carpendale, Sheelagh. "Evaluating information visualizations." Information Visualization. Springer Berlin Heidelberg, 2008. 19-45.
- Heinicke, A., Liao, C., Walbaum, K., Bützler, J., & Schlick, C. M. (2015). User centered evaluation of interactive data visualization forms for document management systems. *Procedia Manufacturing*, 3, 5427-5434.
- Isenberg, Tobias, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Möller. "A Systematic Review on the Practice of Evaluating Visualization." In *IEEE Transactions on Visualization and Computer Graphics (Proc. SciVis 2013)*, 19(12): 2818-2827, 2013.
- Lam, Heidi, et al. "Empirical studies in information visualization: Seven scenarios." *Visualization and Computer Graphics, IEEE Transactions on* 18.9 (2012): 1520-1536.
- Mcgrath, E. "Methodology matters: Doing research in the behavioral and social sciences." *Readings in Human-Computer Interaction: Toward the Year 2000* (2nd ed.), 1995.
- Munzner, Tamara. "A nested model for visualization design and validation." *Visualization and Computer Graphics, IEEE Transactions on* 15.6 (2009): 921-928.
- Plaisant, Catherine. "The challenge of information visualization evaluation." *Proceedings of the working conference on Advanced visual interfaces. ACM*, 2004.
- Saraiya, Purvi, Chris North, and Karen Duca. "An insight-based methodology for evaluating bioinformatics visualizations." *Visualization and Computer Graphics, IEEE Transactions on* 11.4 (2005): 443-456.

Literature

- Shneiderman, Ben, and Catherine Plaisant. "Strategies for evaluating information visualization tools: multi- dimensional in-depth long-term case studies." Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization. ACM, 2006.
- Sedlmair, Michael, Miriah Meyer, and Tamara Munzner. "Design study methodology: Reflections from the trenches and the stacks." Visualization and Computer Graphics, IEEE Transactions on 18.12 (2012): 2431-2440.
- Sedlmair, Michael, et al. "Information visualization evaluation in large companies: Challenges, experiences and recommendations." Information Visualization 10.3 (2011): 248-266.
- South, et al. 2022 - Effective use of Likert scales in visualization evaluations: A systematic review. In Computer Graphics Forum (Vol. 41, No. 3, pp. 43-55).
- Tory, M. (2013). User studies in visualization: A reflection on methods. In Handbook of Human Centric Visualization (pp. 411-426). New York, NY: Springer New York.

Thanks!

Evaluation in Visualization

Guest Lecture
Nov 24, 2017

Michael Sedlmair

slides: [https://homepage.univie.ac.at/michael.sedlmair/
teaching/shandong-eval-2017.pdf](https://homepage.univie.ac.at/michael.sedlmair/teaching/shandong-eval-2017.pdf)