Experimental Design for Visualization Evaluation

Manuela Waldner

Institute of Computer Graphics and Algorithms
Vienna University of Technology, Austria
Why is this important?

- Visualization: „The use of computer-supported, interactive, visual representations of data to amplify cognition.“

- „The purpose of visualization is insight, not pictures. The main goals of insight are discovery, decision making, and explanation.“

[Card et al., Readings in Information Visualization: Using Vision to Think, 1999]
Goals

„We are looking to learn about how information visualizations do or do not support people in their information tasks and/or how people conduct their information related tasks so that visualization can be better designed to support them.“

→ 2 main categories

- Evaluating visualizations
- Understanding data analysis processes

[Carpendale, Evaluating Information Visualizations, Information Visualization, 2008]
Methods

- Controlled experiments / comparative user studies
- Longitudinal field studies
- Inspections performed by experts
  - Heuristic evaluation (informal, holistic)
  - Cognitive walk-throughs (specific tasks)
- Psychophysical evaluations
Which method to use?

Domain problem characterization  Wrong problem!

Data / operation abstraction design  Showing the wrong thing!

Encoding / interaction technique design  Showing it the wrong way!

Algorithm design  Wrong algorithm!

System implementation

[Munzner, A Nested Model for Visualization Design and Validation, InfoVis 2009]
Which method to use?

Domain problem characterization

Validate: interview and observe target users

Data / operation abstraction design

Validate: justify

Encoding / interaction technique design

Validate: field studies

Algorithm design

Validate: analyze computational complexity

System implementation

Validate: user studies

Validate: observe adoption rates

[Munzner, A Nested Model for Visualization Design and Validation, InfoVis 2009]
Controlled User Studies
Controlled User Studies

- **Hypothesis**: logical, precise, testable

- **Independent variables / factors**
  - Examples: visualization method, task, data size, population...
  - Levels: number of variations for each factor = test conditions

- **Dependent variables**: measured user performance
  - Examples: task completion time, accuracy, number of interaction steps, subjective ratings...
Controlled User Studies

Experimental Design:

- **Within-subjects design**: every condition is performed by every participant (= repeated measures)
  - Be aware of learning effect!
- **Between-subjects design**: users are divided into groups, performing different variations of one independent variable
- **Mixed design**: users are divided in groups performing different variations of one independent variable and all variations of others
Controlled User Studies

- **Analysis**
  - How large is the difference?
  - Is the difference statistically significant?
    - t-tests
    - Analysis of Variance (= ANOVA)
  - Is the difference of practical significance?

http://www.socialresearchmethods.net/kb/stat_t.php
Controlled User Studies

- Nested qualitative inquiry: put quantitative data into social context
  - **Experimenter observations** → detect unexpected events, explanations for outliers
  - „think-aloud“
  - **User opinions**
    - Semi-structured interviews
    - Questionnaires (quantification through Likert scale)
Controlled User Studies

Typical tasks for evaluating InfoVis:

- Retrieve a value
- Filter
- Compute derived value (e.g., average, median, count)
- Find extremum
- Sort
- Determine range
- Characterize distribution
- Find anomalies
- Cluster
- Correlate

[Amar et al., Low-Level Components of Analytic Activity in Information Visualization, InfoVis 2005]
Fig. 1. DimpVis is an interaction technique for navigating time in information visualizations through direct manipulation of visualization objects. The hint path reveals the locations of a selected visual item over time. Navigation in time for scatter plots is achieved by dragging a selected point in 2D along its hint path.

[Kondo and Collins, DimpVis: Exploring Time-varying Information Visualizations by Direct Manipulation, TVCG 2014]
Example: DimpVis

Assumption: DimpVis will not be faster or more accurate for reading values, but more efficient for characterizing the trends of data objects.

[Kondo and Collins, DimpVis: Exploring Time-varying Information Visualizations by Direct Manipulation, TVCG 2014]
Example: DimpVis

■ Independent Variables:
  ◆ Technique: DimpVis, Time Slider, Small Multiples
  ◆ Task: retrieve value, compare, characterize distribution, detect outlier (in scatter plots)

■ Dependent Variables:
  ◆ Task completion time (log-transformed)
  ◆ Error rate (incorrect trials / total trials)
  ◆ Subjective ratings (5-point Likert scale)

■ + qualitative, exploratory data (think-aloud)

[Kondo and Collins, DimpVis: Exploring Time-varying Information Visualizations by Direct Manipulation, TVCG 2014]
Example: DimpVis

- **Data:** 20 data points, 10 years
  - Task example: „*When is point A at age=50 and height=5?“*

- **Participants:**
  - 11 males, 2 females
  - Aged 19-30
  - Recruited from „*our university and surrounding area“*

- [Kondo and Collins, DimpVis: Exploring Time-varying Information Visualizations by Direct Manipulation, TVCG 2014]
Example: DimpVis

- **Analysis:**
  - two-way (3x4) repeated measures ANOVA
  - Bonferroni-corrected post-hoc comparisons

- [Kondo and Collins, DimpVis: Exploring Time-varying Information Visualizations by Direct Manipulation, TVCG 2014]
Example: DimpVis

Results:

◆ DimpVis faster than Small Multiples
◆ Small Multiples slowest for outlier detection
◆ Hardly any errors

[Kondo and Collins, DimpVis: Exploring Time-varying Information Visualizations by Direct Manipulation, TVCG 2014]
Example: DimpVis

Discussion

[Kondo and Collins, DimpVis: Exploring Time-varying Information Visualizations by Direct Manipulation, TVCG 2014]
Common challenges:
- Small datasets
- University student participants
- Simple tasks
- State-of-the-art software for comparison
  - Availability
  - Which one to use?
  - Familiarity
- Early prototype
- How to separate error from task completion time?
Controlled User Studies

- Common challenges:
  - Small datasets
  - University student participants
  - Simple tasks
  - State-of-the-art software for comparison
    - Availability
    - Which one to use?
    - Familiarity
  - Early prototype
  - How to separate error from task completion time?
Crowdsourcing Visualization Evaluation

- Web workers perform small tasks for micro-payments
  - Amazon Mechanical Turk

- Advantages:
  - More diverse population $\rightarrow$ ecological validity
  - Large population $\rightarrow$ large design space evaluation

- Disadvantages:
  - Lack of control over experimental conditions
  - Subject motivation?

[Heer and Bostock, Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design, CHI 2010]
Example: Error Bars

- Investigation how different mean and error representations result in differing interpretations of general audience.

Fig. 1. Four encodings for mean and error evaluated in this work. Each prioritizes a different aspect of mean and uncertainty, and results in different patterns of judgment and comprehension for tasks requiring statistical inferences.

[Correll and Gleicher, Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error, TVCG 2014]

Manuela Waldner
Sample hypothesis (there are many more in the paper!):

- Outcomes within the bar of a bar chart will be seen as liklier than outcomes outside of the bar = „within-the-bar“ bias

[Correll and Gleicher, Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error, TVCG 2014]
Example: Error Bars

- **Independent variables:**
  - Encoding (between-subjects)
  - Red dot position above or below mean (within-subjects)

- **Dependent variables:**
  - Perceived likelihood of dot as outcome (7-point Likert scale)

[Correll and Gleicher, Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error, TVCG 2014]
Example: Error Bars

- **Data:** 3 fictional data sets
  - Polling data
  - Weather forecast
  - Financial prediction

- **Participants:**
  - 96 participants = 8 per data set and encoding
  - North American Turker population
  - Balanced gender and education level

[Correll and Gleicher, Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error, TVCG 2014]
Example: Error Bars

Analysis:

- two-way ANCOVA (= analysis of covariance) with Tukey’s Honest Significant Difference (HSD) post-hoc comparisons

- Covariates:
  - inter- and intra-participant variance in performance
  - Data set / problem frame

- Factors = Independent variables
  - 4 encodings
  - 2 dot positions

[Correll and Gleicher, Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error, TVCG 2014]
Example: Error Bars

**Results:**

- Significant interaction between factors
- Only for bar chart: red dots below mean more likely than those above bar

[Correll and Gleicher, Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error, TVCG 2014]
Discussion

[Correll and Gleicher, Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error, TVCG 2014]
Common challenges:
- Small datasets
- University student participants
- **Simple tasks**
- State-of-the-art software for comparison
  - Availability
  - Which one to use?
  - Familiarity
- Early prototype
- How to separate error from task completion time?
Recall:

„The purpose of visualization is **insight**, not pictures. The main goals of insight are **discovery**, **decision making**, and **explanation**.“

[Card et al., Readings in Information Visualization: Using Vision to Think, 1999]

Problem of simple tasks

- Search-like tasks don’t represent insight well!
- Variety of simple tasks does not necessarily generalize to more complex tasks

Insight-based study

- Users explore data until they feel they have learned all they can from the data
- No predefined tasks
- Think-aloud protocol
- Quantification of insights through **coding** method
- Tasks are treated as dependent variables in experiment

Coding

- Iterative characterization of qualitative data (video, field notes,...)

- Open coding:
  - Labeling of each observed phenomenon = concept
  - Grouping of concepts into categories
Example: PathVis

- Analysis of biological experiment:
  - impacts of tabacco smoking on immune response to flu infection

- Pathways (graphs): interactions between bio-molecules (genes, proteins,...)

- Multi-dimensional time series data from gene expression microarrays: expressions of genes over multiple times / conditions

- time series data in context of a graph

[North et al., A comparison of benchmark task and insight evaluation methods for information visualization, Information Visualization 2011]
Example: PathVis

Visualization Method 1: encoding gene expression of one time point in graph (+ time slider)

[North et al., A comparison of benchmark task and insight evaluation methods for information visualization, Information Visualization 2011]
Visualization Method 2: encoding gene expression of multiple time points in single graph

[North et al., A comparison of benchmark task and insight evaluation methods for information visualization, Information Visualization 2011]
Example: PathVis

Visualization Method 3: small multiples showing miniature graphs for each time point

[North et al., A comparison of benchmark task and insight evaluation methods for information visualization, Information Visualization 2011]
Example: PathVis

- **Hypothesis**: each visualization alternative will perform best with tasks that have a query grouping structure that match the visualization’s visual grouping structure

- **Independent variable**: 3 visualization methods

- **Design**: between-subjects

[North et al., A comparison of benchmark task and insight evaluation methods for information visualization, Information Visualization 2011]
Example: PathVis

- **Data:**
  - 46 genes (nodes)
  - 36 gene interactions (edges)
  - 12 gene expressions (time series data)

- **Participants**
  - 30 (between-subjects)
  - Undergraduate students (basic knowledge in molecular biology)

[North et al., A comparison of benchmark task and insight evaluation methods for information visualization, Information Visualization 2011]
Example: PathVis

Analysis:

- ANOVAs of:
  - time spent in study
  - Number of distinct data insights
  - Rate of insights per minute
  - Number of distinct insights for each insight category
- Open coding of think-aloud protocol into insight categories

[North et al., A comparison of benchmark task and insight evaluation methods for information visualization, Information Visualization 2011]
Example: PathVis

- **Results:**
  - Multiple graphs: less time spent in study
  - Single time point: greater number of insights
  - Single time point: fastest insight rate
  - Number of insights correlates with time spent
    → more time spent = more insights

[North et al., A comparison of benchmark task and insight evaluation methods for information visualization, Information Visualization 2011]
Discussion

[North et al., A comparison of benchmark task and insight evaluation methods for information visualization, Information Visualization 2011]
Common challenges:
- Small datasets
- University student participants
- Simple tasks
- State-of-the-art software for comparison
  - Availability
  - Which one to use?
  - Familiarity
- Early prototype
- How to separate error from completion time?
Goal: limit cognitive load associated with interpreting a visualization

Cognitive load measures as DVs
- Subjective data (e.g., NASA-Task Load Index)
- Eye tracking
- Brain activity (EEG)

[Anderson, Evaluating Scientific Visualization Using Cognitive Measures, BELIV 2012]
Eye Tracking

- Recording of spatio-temporal eye movement data
- Raw gaze data $\rightarrow$ filters $\rightarrow$ metrics for statistical analysis:
  - Fixations:
    - Example: number of fixations in an area of interest (AOI)
  - Saccades:
    - Example: saccade amplitude
  - Scanpath (series of fixations and saccades):
    - Example: transition matrix between pairs of AOIs

[Kurzhals et al., Evaluating Visual Analytics with Eye Tracking, BELIV 2014]
Eye Tracking

Visual Analysis

Attention map

Gaze plot

Visual analytics methods from geographic movement data analysis

[Andrienko et al., Visual Analytics Methodology for Eye Movement Studies, TVCG 2012]

[Kurzhals et al., Evaluating Visual Analytics with Eye Tracking, BELIV 2014]
Example: Linear vs. Radial Graphs

Comparison of „retrieve value“ task with four graph types (bar, line, area, scatter) and two graph styles (linear, radial)

[Goldberg and Helfman, Eye tracking for visualization evaluation: Reading values on linear versus radial graphs, Information Visualization 2011]
Example: Linear vs. Radial Graphs

- Areas of Interest (AOIs)

[Goldberg and Helfman, Eye tracking for visualization evaluation: Reading values on linear versus radial graphs, Information Visualization 2011]
Example: Linear vs. Radial Graphs

- **Dependent variables:**
  - Task completion time
  - First fixation time: initial time a fixation was made within an AOI
  - Minimum time: instant a participant had completed initial fixations within all required AOIs
  - ....

- **Analysis:** ANOVA with Tukey’s pairwise comparisons

  [Goldberg and Helfman, Eye tracking for visualization evaluation: Reading values on linear versus radial graphs, Information Visualization 2011]
Example: Linear vs. Radial Graphs

Results:

- Radial graph response time slower
- Mapping datapoint to value (step 3) especially slow for radial graphs

[Goldberg and Helfman, Eye tracking for visualization evaluation: Reading values on linear versus radial graphs, Information Visualization 2011]
Expert Inspections
Heuristic Evaluation

- „discount evaluation method“
  - small set of experts can detect majority of problems
  - Detect positive and negative aspects of a visualization based on set of heuristics

- Information visualization heuristics
  - Example: Shneiderman’s information seeking mantra („overview first, zoom and filter...“)
  - Task-based
  - Perception-based (e.g., „consider people with color blindness“)

[Zuk et al., Heuristics for Information Visualization Evaluation, BELIV 2006]
Target User Observations
Qualitative data acquisition

„Detailed reporting about a small number of individuals working on their own problems, in their normal environment“ [Shneiderman & Plaisant, BELIV 2006]

Unobtrusive, on-site
→ realistic, not generalizable

Observations
- Field notes
- Video or audio tapes
- Logs

Interviews
Example: Cardiogram

- 3-year exploratory study in automotive industry
- Learning about current practices in data analysis of in-car communication networks
- Goal: design of a novel visual analytics tool
  → understanding data analysis processes

[Sedlmair et al., Cardiogram: Visual Analytics for Automotive Engineers, CHI 2011]
Example: Cardiogram

**Method:**

- Multi-dimensional in-depth long-term case study
  - 30 semi-structured interviews
  - 14 contextual inquiries
  - 17 focus groups with 3-10 participants
  - Informal collaboration

[Sedlmair et al., Cardiogram: Visual Analytics for Automotive Engineers, CHI 2011]
Example: Cardiogram

- **Key findings:**
  - Masses of data hard to handle
  - Simple list-based representations do not support temporal analysis, outlier detection, correlations etc.

[Sedlmair et al., Cardiogram: Visual Analytics for Automotive Engineers, CHI 2011]
Other methods with domain experts

- Field experiment
  - On-site „controlled“ study
- Informal evaluation
  - Demo to domain experts in lab
  - No pre-defined task
- User experience /usability studies
  - Qualitative lab study
  - pre-defined task(s)
- Field logs
  - Automatic logging of user behavior in the field
Psychophysical Experiments
Psychophysical experiments

- Methods to measure human sensation triggered by physical stimuli
- Limits of human visual system
  - Finding a flicker frequency so that signal is perceived as steady (absolute threshold)
  - Finding a just noticeable difference (JND) between two colors (difference threshold)
- Precise physical definition of stimulus pattern
- Assuming no / low instructional bias
  → often very low number of participants

Manuela Waldner 62
Psychophysical experiments

- Methods for threshold measurements
  - **Method of adjustment**
    - Observer can adjust stimulus intensity
    - Until it becomes just unnoticeable / just noticeably different from a standard stimulus
  - **Staircase procedure**
    - Dynamic adaptation of stimulus level based on observer’s response
    - Threshold: average stimulus intensities when observer response changed
**Objective:** Evaluation of observers’ sensitivity to local surface orientation

**Independent variables:**
- Surface types (5 different shadings)
- Viewing conditions (mono static, mono motion, stereo static, stereo motion)
- Surface slants (25°, 35°, 45°, 55°)

**Participants:** 4 observers (two paper co-authors)

[Norman et al., The perception of surface orientation from multiple sources of optical information, Perception & Psychophysics 1995]
Method of adjustment:
- Orienting a probe towards the estimated surface normal

[Norman et al., The perception of surface orientation from multiple sources of optical information, Perception & Psychophysics 1995]
Example: Surface Orientation Perception

- **Analysis and Results:**
  - Repeated measures ANOVA on angular difference: mono+static largest error
  - Analysis of linear regression on slant: low correlation coefficient $r^2$

[Norman et al., The perception of surface orientation from multiple sources of optical information, Perception & Psychophysics 1995]

Manuela Waldner
Final Example:
InfoVis Best Paper 2014
Example: DOSA

Multivariate network exploration by going from detail to overview via selections and aggregations

[van den Elzen and van Wijk, Multivariate Network Exploration and Presentation: From Detail to Overview via Selections and Aggregations, TVCG 2014]
Related work: either focused on structure or multivariate data exploration

[Perer and Shneiderman, Balancing Systematic and Flexible Exploration of Social Networks, TVCG 2006]

[Bezerianos et al., GraphDice: A System for Exploring Multivariate Social Networks, EuroVis 2010]
Example: DOSA

- **Hypothesis**: the greatest insights from multivariate networks are gained from simultaneous exploration of structure and multivariate data
- **Target audience**: both, experts and casual users

[van den Elzen and van Wijk, Multivariate Network Exploration and Presentation:From Detail to Overview via Selections and Aggregations, TVCG 2014]
Example: DOSA

How to evaluate?
Example: DOSA

2 Use Cases

- US migration and census
- Company e-mail traffic

[van den Elzen and van Wijk, Multivariate Network Exploration and Presentation: From Detail to Overview via Selections and Aggregations, TVCG 2014]
Example: DOSA

- **US Migration and Census**
  - Nodes: 3321 counties
  - Edges: 78294 migrations
  - 14 node attributes
  - 10 edge attributes

[van den Elzen and van Wijk, Multivariate Network Exploration and Presentation:From Detail to Overview via Selections and Aggregations, TVCG 2014]
Example: DOSA

- US Migration and Census
  - Node layout according to longitude / latitude

[van den Elzen and van Wijk, Multivariate Network Exploration and Presentation: From Detail to Overview via Selections and Aggregations, TVCG 2014]
Example: DOSA

- US Migration and Census
  - Selection around New York area with outbound migration

[van den Elzen and van Wijk, Multivariate Network Exploration and Presentation: From Detail to Overview via Selections and Aggregations, TVCG 2014]
Example: DOSA

- US Migration and Census
  - Multiple age selections

People from low-age counties tend to migrate to other low-age counties

[van den Elzen and van Wijk, Multivariate Network Exploration and Presentation: From Detail to Overview via Selections and Aggregations, TVCG 2014]

Manuela Waldner
Example: DOSA

- Discussion
Qualitative Result Inspection

- No experiment
- Reader inspects result image or walkthrough
  - Comparative
  - Isolated

[Isenberg et al., A Systematic Review on the Practice of Evaluating Visualization, TVCG 2013]
Summary
Different methods for different design levels

- „Classic“ controlled user studies
- Alternative experimental designs:
  - Crowd sourcing
  - Insight-based evaluation
  - Eye tracking
- Heuristic Evaluation
- Field Studies
- Psychophysical experiments
- Alternative to user studies: qualitative result inspections
## Methods Overview

<table>
<thead>
<tr>
<th>Method</th>
<th>Sample</th>
<th>Duration</th>
<th>Environment</th>
<th>Data</th>
<th>Task</th>
<th>Realism</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controlled User Study</td>
<td>Large, usually non-experts</td>
<td>Short</td>
<td>Lab</td>
<td>Quantitative (+ nested qualitative data)</td>
<td>Simple and short</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Usability Study</td>
<td>Small, domain experts</td>
<td>Short</td>
<td>Lab</td>
<td>Mostly qualitative</td>
<td>Medium complexity</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Heuristic evaluation</td>
<td>Small, visualization experts</td>
<td>Short</td>
<td>Lab</td>
<td>Quantitative + qualitative</td>
<td>Medium complexity</td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td>Informal evaluation</td>
<td>Small, domain experts</td>
<td>Short</td>
<td>Lab</td>
<td>Qualitative</td>
<td>---</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Field experiment</td>
<td>Small, domain experts</td>
<td>Short</td>
<td>Field</td>
<td>Mostly qualitative</td>
<td>Medium complexity</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Case / field study</td>
<td>Small, domain experts</td>
<td>Long</td>
<td>Field</td>
<td>Mostly qualitative</td>
<td>---</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Field logs</td>
<td>Large, domain experts</td>
<td>Long</td>
<td>Field</td>
<td>Mostly quantitative</td>
<td>---</td>
<td>High</td>
<td>Medium</td>
</tr>
</tbody>
</table>
General References

- Munzner, *A nested model for visualization design and validation*, TVCG 2009
- Lam et al., *Empirical Studies in Information Visualization: Seven Scenarios*, TVCG 2012
- Isenberg et al., *A Systematic Review on the Practie of Evaluating Visualization*, TVCG 2013
- Ellis and Dix, *An explorative analysis of user evaluation studies in information visualisation*, BELIV 2006
- Ware, *Information Visualization: Perception for Design*, 2013 (3rd edition)
- Ehrenstein and Ehrenstein, *Psychophysical Methods*, Modern Techniques in Neuroscience Research, 1999