Readings Covered

Ware, Appendix C: The Perceptual Evaluation of Visualization Techniques and Systems


Further Readings

Task-Centered User Interface Design, Clayton Lewis and John Rieman, Chapters 0-5.

Ware: Evaluation Appendix

- perceptual evaluation of infovis techniques and systems
  - empirical research methods applied to vis
  - difficult to isolate evaluation to perception
- research method depends on research question and object under study

[Ware, Appendix C: The Perceptual Evaluation of Visualization Techniques and Systems. Information Visualization: Perception for Design. ]
Psychophysics

- **method of limits**
  - find limitations of human perceptions
- **error detection methods**
  - find threshold of performance degradation
  - staircase procedure to find threshold faster
- **method of adjustment**
  - find optimal level of stimuli by letting subjects control the level
Cognitive Psychology

- repeating simple, but important tasks, and measure reaction time or error
  - Miller’s 7+/- 2 short-term memory experiments
  - Fitts’ Law (target selection)
  - Hick’s Law (decision making given n choices)
- interference between channels
- multi-modal studies
  - using haptic feedback for interruption when the participants were visually (and cognitively) busy
Structural Analysis

- requirement analysis, task analysis
- structured interviews
  - can be used almost anywhere, for open-ended questions and answers
- rating/Likert scales
  - commonly used to solicit subjective feedback
  - ex: NASA-TLX (Task Load Index) to assess mental workload
    - “it is frustrating to use the interface”
    - Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree
Comparative User Studies

- hypothesis testing

- hypothesis: a precise problem statement
  - ex from Snap: Participants will be faster with a coordinated overview+detail display than with an uncoordinated display or a detail-only display with the task requires reading details

- measurement: faster
- objects of comparison:
  - coordinated O+D display
  - uncoordinated O display
  - uncoordinated D display

- condition of comparison: task requires reading details
Comparative User Studies

- study design: factors and levels

- factors
  - independent variables
  - ex: interface, task, participant demographics

- levels
  - number of variables in each factor
  - limited by length of study and number of participants
Comparative User Studies

- study design: within, or between?
- within
  - everybody does all the conditions
  - can lead to ordering effects
  - can account for individual differences and reduce noise
  - thus can be more powerful and require fewer participants
  - combinatorial explosion
    - severe limits on number of conditions
  - possible workaround is multiple sessions
- between
  - divide participants into groups
  - each group does only some conditions
Comparative User Studies

- measurements (dependent variables)
  - performance indicators: task completion time, error rates, mouse movement
  - subjective participant feedback: satisfaction ratings, closed-ended questions, interview
  - observations: behaviors, signs of frustration

- number of participants
  - depends on effect size and study design: power of experiment

- possible confounds?
  - learning effect: did everybody use interfaces in a certain order?
  - if so, are people faster because they are more practiced, or because of true interface effect?
Comparative User Studies

- result analysis
  - should know how to analyze the main results/hypotheses BEFORE study
  - hypothesis testing analysis (using ANOVA or t-tests) tests how likely observed differences between groups are due to chance alone
    - ex: a p-value of 0.05 means there is a 5% probability the difference occurred by chance
      - usually good enough for HCI studies

- pilots!
  - should know the main results of the study BEFORE actual study
Evalation Throughout Design Cycle

- user/task centered design cycle
  - initial assessments
  - iterative design process
  - benchmarking
  - deployment

- identify problems, go back to previous step

Task-Centered User Interface Design, Clayton Lewis and John Rieman, Chapters 0-5.
Initial Assessments

- what kind of problems are the system aiming to address?
  - analyze a large and complex dataset
- who are your target users?
  - data analysts
- what are the tasks? what are the goals?
  - find trends and patterns in the data via exploratory analysis
- what are their current practices
  - statistical analysis
- why and how can visualization be useful?
  - visual spotting of trends and patterns
- talk to the users, and observe what they do
- task analysis
Iterative Design Process

- does your design address the users’ needs?
- can they use it?
- where are the usability problems?

- evaluate without users
  - cognitive walkthrough
  - action analysis
  - heuristics analysis

- evaluate with users
  - usability evaluations (think-aloud)
  - bottom-line measurements
  - example: snap paper experiment 1
Benchmarking

- how does your system compare to existing ones?
  - snap paper experiment 2
- empirical, comparative studies
  - ask specific questions
  - compare an aspect of the system with specific tasks
    - Amar/Stasko task taxonomy paper
- quantitative, but limited
Deployment

- how is the system used in the wild?
- how are people using it?
- does the system fit into existing work flow? environment?
- contextual studies, field studies
Comparing Systems vs. Characterizing Usage

- user/task centered design cycle:
  - initial assessments
  - iterative design process
  - benchmarking: head-to-head comparison
  - deployment
  - (identify problems, go back to previous step)
- understanding/characterizing techniques
  - tease apart factors
  - when and how is technique appropriate
- line is blurry: intent
Snap-Together Visualization: Can Users Construct and Operate Coordinated Views?
Snap CMV Formalism

- relation :: visualization
- tuple :: item
- primary key :: item ID
- join :: coordination

Snap CMV Formalism

- one-to-one
  - linked selection across views
  - overview select → detail scroll
  - linked scrolling across views

- one-to-many
  - parent select → child load

- architecture
  - independent modules linked via API
  - versus tightly coupled Improvise approach
Snap Usability Evaluation

- 6 participants: 3 data analysts, 3 programmers
  - census bureau: analysts + 1 programmer (expert?)
  - CS students: 2 programmers (novice?)
- 3 tasks
  - 2 construct to spec
  - 1 open ended, “abstract thinking about coordination”
- Measurements
  - survey of background knowledge (data, tools)
  - success at task
  - learning time, time to completion
Snap Usability Results

- success, enthusiasm
  - possible confound from please-the-creator effect
- analyst/programmer differences
  - interface building as exploration vs. construction
  - analysts performed better
- snap usability problems
  - explicit overview of coordination setup may help
  - provide attribute lists instead of requiring access queries
  - window rearrangement timesink
Snap User Study

- **hypothesis**
  - participants will be faster with a coordinated overview+detail display than with an uncoordinated display or a detail-only display with the task requires reading details

- **factors and levels**
  - **interface**: 3 levels
    - detail-only
    - uncoordinated overview+detail
    - coordinated overview+detail
  - **task**: 9 levels
    - many browsing tasks, not grouped prior to study
    - closed-ended, with obvious correct answers
    - ex: “which state has the highest college degree
    - compare with open-ended usability task: “Please create a user-interface that will support users in efficiently performing the following task: to be able to quickly discover which states have high population and high Per Capita Income, and examine their counties with the most employees”
Snap User Study Design

- within-subject
  - everybody worked on all interfaces/task combos
- counterbalanced between interfaces
  - 6 permutations to avoid ordering / learning effects
    - 3 groups x 6 permutations = 18 participants
- need one task set (9) for each interface
  - tasks in each set need to be isomorphic
- 27 tasks per study per participant
  - 3 interfaces x 9 tasks
Snap User Study Design

- **measurements**
  - task completion time to obtain answer
    - no errors
  - subjective ratings using rated scale (1-9)
- **participants**
  - 18 students (novice)
Snap User Study

- time result analysis: hypothesis testing with ANOVA
  - 3 (interface) x 9 (task) within-subjects ANOVA to check for main effects of interface, or task, or interface/task interaction

- ANOVA
  - (ANalysis Of VAriance between groups)
  - commonly used statistics for factorial designs
  - tests difference between means of two or more groups
    - example use: two-way ANOVA to see if there is an effect of interface and task, or interaction between them
Snap User Study

- time result analysis: descriptive statistics
  - on average, coordination achieves an 80% speedup over detail-only for all tasks
  - good for discoveries based on results
  - example: 3 task groups
  - example: explain quantitative data with observed participant behaviours

- subjective satisfaction analysis: hypothesis testing with ANOVA
  - 3 (interface) x 4 (question category) within-subjects ANOVA
Critique

- good example of usability vs. comparative study

<table>
<thead>
<tr>
<th></th>
<th>Usability testing</th>
<th>User study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aim</td>
<td>Improve product design</td>
<td>• Discover knowledge</td>
</tr>
<tr>
<td></td>
<td>• Is the prototype usable?</td>
<td>(how are interfaces used?)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Prove concepts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Is your novel technique actually useful?)</td>
</tr>
<tr>
<td>Participants</td>
<td>Few, domain expert or target users</td>
<td>More, novices, general human behaviours</td>
</tr>
<tr>
<td>Expt conditions</td>
<td>Partially controlled, and could be realistic,</td>
<td>Strongly controlled, unrealistic</td>
</tr>
<tr>
<td></td>
<td>more open-ended tasks</td>
<td>laboratory environment with predefined,</td>
</tr>
<tr>
<td></td>
<td>◊ More ecologically valid?</td>
<td>simplistic tasks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>◊ Less ecologically valid?</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>Not perfectly replicable, too many uncontrolled /</td>
<td>Should be replicable</td>
</tr>
<tr>
<td></td>
<td>uncontrollable factors</td>
<td>(but, limited generalizbility?)</td>
</tr>
<tr>
<td>Report to...</td>
<td>Developers</td>
<td>Scientific community</td>
</tr>
<tr>
<td>Bottom-line</td>
<td>Identify usability problems</td>
<td>Hypothesis testing (yes, need those p-values</td>
</tr>
<tr>
<td></td>
<td></td>
<td>to be less than .05!)</td>
</tr>
</tbody>
</table>

Perceptual Scalability

- what are perceptual/cognitive limits when screen-space constraints lifted?
  - 2 vs. 32 Mpixel display
  - macro/micro views
- perceptually scalable
  - no increase in task completion times when normalize to amount of data

Perceptual Scalability

- design
  - 2 display sizes, between-subjects
    - (data size also increased proportionally)
  - 3 visualization designs, within
    - small multiples: bars
    - embedded graphs
    - embedded bars
  - 7 tasks, within
  - 42 tasks per participant
    - 3 vis x 7 tasks x 2 trials
Embedded Visualizations

Small Multiples Visualizations

- attribute-centric instead of space-centric

Results

- 20x increase in data, but only 3x increase in absolute task times

Results

- significant 3-way interaction
  - between display, size, task

Results

- visual encoding important on small displays
  - DS: mults sig slower than graphs on small
  - DS: mults sig slower than embedded on large
  - OS: bars sig faster than graphs for small
  - OS: no sig difference bars/graphs for large

- spatial grouping important on large displays
  - embedded sig faster+preferred over small mult
  - no bar/graph differences
Critique

- first study of macro/micro effects
  - breaking new ground

- many possible followups
  - physical navigation vs. virtual navigation
Fisheye Multilevel Networks

[Navigating Hierarchically Clustered Networks through Fisheye and Full-Zoom Methods. Schaffer et al. ACM ToCHI 3(2) p 162-188, 1996.]
Lab Experiment

- 2 interfaces (fisheye, zoom)
- 2 tasks (isomorphic)
  - stages: find and repair
- within subjects, counterbalanced order
- 20 participants
- data: 154 nodes, 39 clusters
- measurements
  - completion time
  - number of zooms
  - success
Results

- sig effect of interface: fisheye faster
- but no differences with find subtask
  - information visible in both displays
- solution quality differed: fisheye better
  - local rerouting difficult in full-zoom
Field Experiment

- 2 real control room operators
- response times similar
  - no statistical analysis, too few subjects
- expressed preference for fisheye over full-zoom
  - (experimenter effect?)
- concerns about fisheye: missing details
Critique

- nicely designed study
- useful discussion of qualitative observations
- very good to do field followup with real operators
Pictures Into Numbers

- field study
- participants: professional meterologists
  - two people: forecaster, technician
- interfaces: multiple programs used
- protocol
  - talkaloud
  - videotaped sessions with 3 cameras

Cognitive Task Analysis

- initialize understanding of large scale weather
- build qualitative mental model (QMM)
- verify and adjust QMM
- write the brief
- task breakdown part of paper contribution
Coding Methodology

- interface
  - which interface used
  - whether picture/chart/graph
- usage (every utterance!)
  - goal
  - extract
    - quant/qual
    - goal-oriented/opportunistic
    - integrated/unintegrated
- brief-writing
  - quant/qual
  - QMM/vis/notes
Results

- sig difference between vis used at CTA stages
  - charts to build QMM
  - images to verify/adjust QMM
  - all kinds during brief-writing

- many others...

Critique

- video coding is huge amount of work, but very illuminating
  - untangling complex story of real tool use
- methodology of CTA construction not discussed here
  - often bottomup/topdown mix
Credits

- Heidi Lam guest lecture

http://www.cs.ubc.ca/ tmm/courses/cpsc533c-06-fall/#lect10