CPSC 533C
Evaluation

Heidi Lam

Oct 12, 2006

Readings

Readings

- The Perceptual Evaluation of Visualization Techniques and Systems. Ware, Appendix C.
- Snap-Together Visualization: Can Users Construct and Operate Coordinated Views? North, C. and Shneiderman B. *Intl. Journal of Human-Computer Studies* 53(5), p. 715-739, 2000.
- Low-Level Components of Analytic Activity in Information Visualization. Amar, R., Eagan, J. and Stasko, J. In *Proc InfoVis*, p. 111-117, 2005.

Further Readings

- Task-Centered User Interface Design. Chapters 0-5. Lewis, C. and Rieman, J.
- The Challenge of Information Visualization Evaluation. Plaisant, C. In *Proc Advanced Visual Interfaces (AVI)*, 2004.

Interface Design and Evaluation

Evaluation is required at all stages in system development

1. Initial assessments:

   - What kind of problems are the system aiming to address?
     (e.g., difficult to analyze a large and complex dataset)
   - Who is your target users?
     (e.g., data analysts)
   - What are the tasks? What are the goals?
     (e.g., to find trends and patterns in the data via exploratory analysis)
   - What are their current practice?
     (e.g., statistical analysis)
   - Why and how can visualization be useful?
     (e.g., visual spotting of trends and patterns)

   - Talk to the users, and observe what they do
   - Task analysis

Interface Design and Evaluation

Evaluation is required at all stages in system development

1. Initial assessments
2. 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
  (e.g., the snap-together paper experiment 1)

Interface Design and Evaluation

Evaluation is required at all stages in system development

1. Initial assessment
2. Iterative design process
3. Bench-marking:

- How does your system compare to existing systems?
  (e.g., the Snap-together paper, experiment 2)

- Empirical, comparative user studies
  - Ask specific questions
  - Compare an aspect of the system with specific tasks
    (task taxonomy paper; Ware's appendix C)
  - Quantitative, but limited
    (see *The Challenge of Information Visualization Evaluation*)

Interface Design and Evaluation

Evaluation is required at all stages in system development

1. Initial assessments
2. Iterative design process
3. Bench-marking
4. Deployment:
   - How is the system used in the wild?
   - Are people using it?
   - Does the system fit in with existing work flow? Environment?

   Contextual studies, field studies…

Interface Design and Evaluation

Evaluation is required at all stages in system development

1. Initial assessments
2. Iterative design process
3. Bench-marking
4. Deployment
5. Identify problems and go back to 1, 2, 3, or 4

Snap-Together Visualization: Can Users Construct and Operate Coordinated Views?

North and Shneiderman, 2000

Usability Evaluation

Snap-Together Visualization: usability evaluation

- Goal
  - To evaluate the usability and benefit of the Snap system itself and discover potential user-interface improvements

- Participants
  - 3 data analysts--familiar with data and analysis as they were employees of the US Bureau of the Census and the study used census data
  - 3 programmers--1 from the Census, and 2 CS students on campus
  - Domain experts vs. novices? Part of the design?

Snap-Together Visualization: usability evaluation

- Tasks
  - 3 exercises to construct a coordinated-visualization user interface according to a provided specification
  - Exercises designed to test different aspects of the system to uncover usability issues
  - First 2 exercises were interface construction according to spec (screenshots); Exercise 3 was more open-ended that required "abstract thinking about coordination, task-oriented user-interface design".
  - Did not say how these tasks were chosen. For example, is "one-to-many" join relationship (Exercise 2) suspected to be difficult prior to the study?

Snap-Together Visualization: usability evaluation

- Procedures:
  - Did not say if participants think aloud (so, how did the experimenter identify "cognitive trouble spots in training and test trials, and Snap user-interface problems"?)
- Measurements:
  - Subjects' background information from a survey, on experience on Access / SQL, and on the data
  - Success
  - Learning time, and time to completion
- Observations:
  - Cognitive trouble spots
  - Snap user-interface problems

Snap-Together Visualization: usability evaluation

- Results:
  - Timing Results: hard to interpret (no bottom-line)
    - Is it ok to spend 10-15 minutes on Exercise 3?|
  - Success: also hard to interpret as did not report in what form and how frequently the help was provided
  - Reported differences between analysts and programmers
    - Analysts considered interface building as exploration; programmers as construction
    - Analysts performed better
    - Would be more useful to identify individuals in their report (Customary to say CS student 1 did this, Analysts 1 did that…)
    - For example, did the Access/SQL experience of the Census programmer made a difference?

Snap-Together Visualization: usability evaluation

- Results:
  - Qualitative observations were vague, and with possible confound
    - "In general, the subjects were quick to learn the concepts and usage, and were very capable to construct their own coordinated-visualization interfaces"
    - "There may have been social pressure to respond positively, since the subjects knew that the administrator of the experiment was also the developer of the Snap system"
  - Identified 4 usability problems
    - Should probably rate the severity of the problems
    - Not sure if they changed Snap before the second study
    - Did not close the loop by re-evaluation

Your Questions: about the snap idea

- One thing that struck me about this paper is that it appears to give more credit to intense user interaction than is really needed. Firstly, the paper gives off the impression that users are "constructing" and "linking" visualizations from ether when in fact much of what can be done (multiple views, linked visualizations) is already pre-determined for them in the sense that they are merely asking to visualize things in different pre-defined visualization types. Additionally, many of these defined vis types are rather broad and could possibly fail to address context-specific data.

  In fact, the extra work users have to do in setting up links etc. does not appear to give much more benefit than a context-specific, pre-constructed visualization system that offers multiple linked views of the same data. Besides, if they knew enough about the domain to "construct" a visualization around it, then they already knew what they were looking for!

Your Questions: about the snap idea

- In the "Snap" system, user needs to drag-and-drop snap button to another window to coordinate visualizations but is not it the whole idea of the system to make coordinated visualizations? The dialog should pop-up automatically when another method of data representation is going to be shown or even provide default coordination with ability to edit it.

- Coordination when visualizing multidimensional data could appear pretty difficult task since many visual dimensions are represented by certain techniques, highlighting which can cause loss of the visual perception of a data. How to address that?

- There is also a problem that can appear with uncertainty of scrolling in case of the list as one representation and, for instance, a focus+context as a second coordinated representation. When we are scrolling the list, should we jump, chaotically changing focus, in the other visual representation or should we just highlight the position of the object? If we choose the second case we are risking not to find it on the screen at all because of the size on the edges of the distortion but if we choose the first case then we easily lose track of the position where we are located in the data globally.

Your Questions: Usability Evaluation

- In the first evaluation experiment, I noticed there was no control group. Maybe their evaluation was just to check that there was noting really bad with their idea. However, if they wanted to see any more than that, I think they should have compared against at least a couple of people that were using standard tools. They say that window management is a problem, taking a lot of time. It would be interesting and important to check that any time savings as a result of the re-organized windows aren't offset by the time it takes to set up the windows, especially for infrequent tasks.

The Perceptual Evaluation of Visualization Techniques and Systems

Ware

The Perceptual Evaluation of Visualization Techniques and Systems

- More like: empirical research methods applied to visualization, as it is oftentimes difficult to isolate the evaluation to perception

- The research method selected depends on the research question and the object under study

- Will not cover some of the methods in the appendix that are for data analysis (e.g., the Statistical Exploration section), and some that are specialized topics (Cross-cultural studies and Child studies)

The Perceptual Evaluation of Visualization Techniques and Systems

- Psychophysics
  - Method of Limits: Find limitations of human perceptions
    - E.g., work from The Sensory Perception and Interaction Research Group of Dr. Karon MacLean:
      finding building blocks of haptic communication as in "Haptic Phoneme", or the smallest unit of a constructed haptic signal to which a meaning can be assigned
  - Error detection methods: Find threshold of performance degradation
    - E.g., Dr. Ron Rensink et al. conducted an experiment to identify the effect of adding visual flow in car-speed judgment that used the staircase procedure to capture thresholds
  - Method of Adjustment: Find optimal level of stimuli by letting subjects control the level

The Perceptual Evaluation of Visualization Techniques and Systems

- Cognitive Psychology
  - Repeating simple, but important tasks, and measure reaction time or error
    - E.g., Miller's 7+/- 2 short-term memory experiments
    - Fitt's Law (target selection)
    - Hick's Law (decision making given *n* choices)

      …
  - Multi-modal studies
    - E.g., MacLean's SPIN lab work "Perceiving Ordinal Data Haptically Under Workload", 2005, using haptic feedback for interruption when the participants were visually (and cognitively) busy

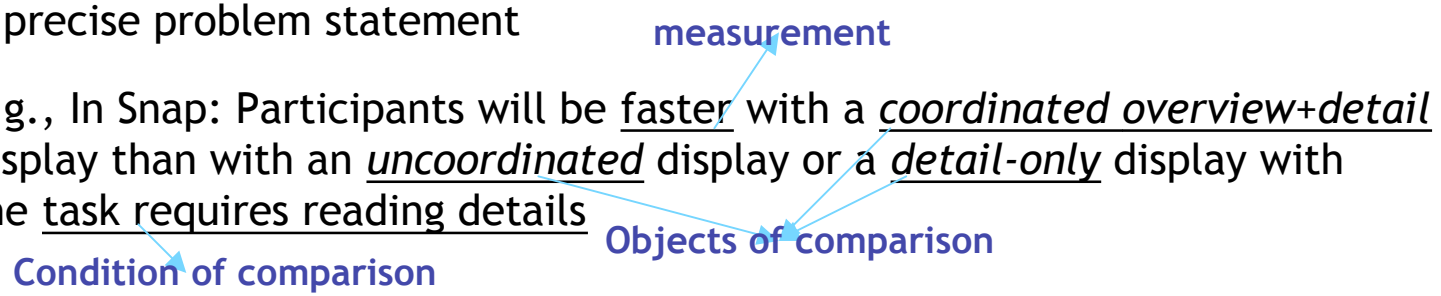
The Perceptual Evaluation of Visualization Techniques and Systems

- Structural Analysis
  - Requirement analysis, task analysis
  - Structured interviews
    - Can be used almost anywhere, for open-ended questions and answers
  - Rating Scales
    - Commonly used to solicit subjective feedback
    - E.g., NASA-TLX (Task Load Index) to assess mental workload
    - E.g., "It is frustrating to use the interface"
      - Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree

The Perceptual Evaluation of Visualization Techniques and Systems

Comparative user study: Hypothesis testing

- Hypothesis
  - A precise problem statement
  - E.g., In 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



- Factors
  - Independent variables
  - E.g., interface, task, participant demographics…
- Levels
  - The number of variables in each factor
  - Limited by the length of the study and the number of participants

The Perceptual Evaluation of Visualization Techniques and Systems

Comparative user study

- Study design: Within, or between?
  - Within
    - Everybody does all the conditions (interface A, task 1…9; interface B, task 1…9, interface C, task 1…9)
    - Can account for individual differences and reduce noise (that's why it may be more powerful and requires less participants)
    - Severely limits the number of conditions, and even types of tasks tested (*may be* able to workaround by having multiple sessions)
    - Can lead to ordering effects
  - Between
    - Divide the participants into group, and each group does some of the conditions

The Perceptual Evaluation of Visualization Techniques and Systems

Comparative user study

- Measurements (dependent variables)
  - Performance indicators: task completion time, error rates, mouse movement…
  - Subjective participant feedback: satisfaction ratings, closed-ended questions, interviews…
  - Observations: behaviors, signs of frustrations…
- Number of participants
  - Depends on effect size and study design--power of experiment
- Possible confounds?
  - Learning effect: Did everybody use the interface in a certain order? If so, are people faster because they are more practiced, or because of the effect of the interface?

The Perceptual Evaluation of Visualization Techniques and Systems

Comparative user study

- Result analysis
  - Should know how to analyze the main results/hypotheses **BEFORE** the study
  - Hypothesis testing analysis using ANOVA or t-test tests how likely observed differences between groups are due to chance alone. For example, a *p*-value of 0.05 means, "there is a 5% probability the difference occurred by chance", which is usually good enough for HCI studies.
- Pilots!
  - Should know the main results of the study **BEFORE** the actual study

Your Questions: Evaluation in practice

- How much work in information visualization is actually informed by psychophysics and cognitive psychology? Aren't the design decisions generally at a much higher level, and based on user studies or even just what seems like a good idea?

- Ware talks about evaluation of systems within the research field. Is there similar focus on evaluation or verification of production visualization systems? Do the standard software engineering principles apply?

- There is a part about bias in "The Perceptual Evaluation of Visualization Techniques and Systems" that tells how important to avoid bias that can change user perception. But could bias positively influence a user judgment? For example, there is a new visualization of some data in which we want to find some patterns. We cannot know if they exist but if we tell to the analyst that patterns must be there the analyst might find them because of the greater determination in the search (of course there is a probability of the mislead).

Your Questions: Evaluation in practice

- Ware suggests using PCA or other dimensional reduction methods to determine how many dimensions are *really* needed based on subjects' responses to the different dimensions they perceive. He states that in reducing dimensionality we can see the actual number of dimensions we need to represent the data. However, is it necessarily true that the minimal number of dimensions to represent the data = the best choice? While cutting down on dimensions helps our perceptual mechanisms directly, it also can potentially make the resultant transformed dimensions increasingly abstract relative to our understanding of the domain, and may not correspond to what a typical user sees as being a useful dimension in terms of what she wants out of the data.

- Ware calls the practice of comparing a display method to a poor alternative dishonest, but isn't there a value in comparing some technique (e.g. edge lens) with the most basic alternative (e.g. a plain node-link diagram with no interactivity)? A comparison with the best current method is of course valuable, but what if the technique being investigated is not compatible with it? In that case, a different path is being explored, so it should also be compared with the starting point. Even if a new technique by itself is not better than the current best practice, perhaps further extensions of it will be.

Your Questions: Empirical Evaluation

- The appendix C of Ware's book introduces several methods relating to the perceptual evaluation of visualization techniques and system. However, it did not elaborate frameworks utilizing mentioned methods in evaluating visualization in terms of system level performance, specific techniques performance, and low-level visual effects performance respectively. Maybe it is just too difficult to come up with a complete evaluation framework if we examine the issue of "combinatorial explosion" raised by the appendix. There are just too many variables, and interaction amongst them, needed to be considered when evaluating a visualization system. In the end, it will become an evaluation task of answering why most of the people use Microsoft's Window over Mac. OS interface when the visualization system gets more complex.

- I think the most used and feasible method of evaluating a visualization system is to compare the times and correctness of performing information tasks with other different visualization systems. Besides the evaluation of system performance, it is also crucial to evaluate which visualization elements or techniques contribute to the enhancement of the task performance and how or what are their contributions. Although it is a good way to refine any system, it could be tedious, costly and beyond the capability of academia.

- Ware mentions about statistical consulting services in many universalities. Do we have one of those? What about ethics? What is the process of submitting an ethics approval?

Low-Level Components of Analytic Activity in Information Visualization.

Amar, Eagan, and Stasko

Low-Level Components of Analytic Activity in Information Visualization

- How to select tasks for a user study?
- Generally, use tasks that the interface is designed for
  - Can directly see if the design is successful over competitor
  - But, hard for researchers to see if the new visualization technique is useful elsewhere
  - Need a standardized task metrics for comparison
  - BUT, the tasks are atomic and simple, may not reflect real-world tasks

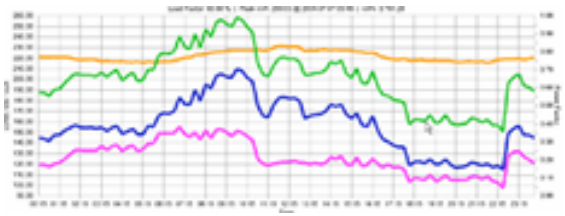
Low-Level Components of Analytic Activity in Information Visualization

- Identified 10 low-level analysis tasks that largely capture people's activities while employing information visualization tools for understanding data

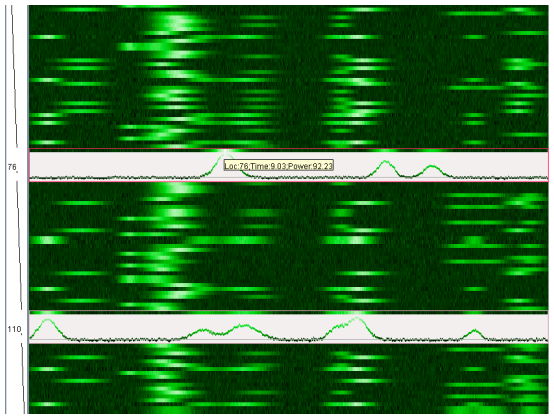
- Retrieve value
- Filter
- Compute derived value
- Find extremum
- Sort
- Determine range
- Characterize distribution
- Find anomalies
- Cluster
- Correlate

Low-Level Components of Analytic Activity in Information Visualization

- We could study tasks based on these operations



Real power consumption data



One possible study interfaces

- E.g. find extremum: Find data cases possessing an extreme value of an attribute over its range within the data set (Amar, 2005)
- In the scenario of monitoring and managing electric power in a control room…
- Which location has the highest power surge for the given time period? (extreme y-dimension)
- A fault occurred at the beginning of this recording, and resulted in a temporary power surge. Which location is affected the earliest? (extreme x-dimension)
- Which location has the most number of power surges? (extreme count)

Your Questions: the task identification approach

- Their analysis generates much insight, but it pre-supposes a numerical / computational specification to the nature of low-level tasks. While it is true that data sets are ultimately represented as numbers in a visualization's backend, it is not necessarily true that the data itself is inherently 'numerical' in the abstract sense (we do not think of maps as sets of numbers, for instance, but as geographical regions). The authors seem to take the point of view that we should go from abstract data->numbers->visualization when performing a task, although from a task-based point of view we should really be looking at how the visualizations represent abstract information directly (especially in the case of maps, which are not inherently numeric).

  This is further reflected in the fact that they gave the users the visualizations to use BEFORE having them look at the data; shouldn't they ask the users *what* they would like to look for in the data, a priori, before presenting them with a particular ideology encapsulated in a visualization? It becomes almost circular to reinforce current task-based notions of visualizations if the root causes of why visualizations are needed to support certain tasks are not addressed.

- In the section on methodological concerns, they point out several sources of bias, but do not point out the directions that were given to the students. At the end of their directions there was a statement instructing them to recall "identification, outliers, correlations, clusters, trends, associations, presentation, etc.". I think that this questions would have caused as much if not more bias as all of the other factors mentioned by the authors

Your Questions: taxonomy use

- The taxonomy seems like it would be worth thinking about when designing an application, but shouldn't be considered a goal or a measure of the utility of an application. Different applications require different sets of tasks, and it may be detrimental to attempt to formulate every task as a composition of primitive tasks. Part of the power of visualizations is that they can reduce the need for users to construct database queries, but this taxonomy seems to be moving the other way, trying to construct visualization tasks in terms of queries.

- The authors motivate this study by giving the example that insights generated from tools used to visualize gene expression data were not generally valuable to domain experts. Then, they limit their study to low-level inquiries. What if it was exactly these low level inquiries that the experts in the gene domain just didn't need?

Your Questions: taxonomy use

- The low-level analytic tasks are like the fundamental construction methods, data is like materials, and the visualization can be seen as state-of-the-art tools that facilitate the construction process. So there is another issue of how to dismantle a general informational question (product) into combinations of low-level tasks and data required (methods and materials), and how to select adequate visualization that can facilitate the tasks. Although it needs practice and training, this approach provides users a systematic approach for using visualization systems.

- The "Low-Level Components of Analytic Activity in Information Visualization" tells about automatic choice of presentation. The paper does not state it is based on data analysis. They are kind of separate but should be together. It would be good first to find automatically "interesting" points about a data and then automatically choose the visualization for good interpretation. What kind of effective data mining algorithms exist for the searching of data internal relations, deviations, correlations and so on besides Bayesian filters applied to the data?

Your Questions

- Was qualitative questions such as, "which system did you like better?" intentionally left out because that is difficult for such vague answers to "serve as a form of common language or vocabulary when discussing the capabilities, advantages, and weaknesses of different information visualization systems."

- Can this paper be extended to general HCI? Is there a taxonomy of questions the general HCI?

- After reviewing this paper, I have a refreshing thought regarding the so called "data analysis". I think the term of "data analysis" is the products of statistics. Since the information visualization can not provide the rigorous analysis results like the statistical methods do, we might just shift to ask meaningful questions rather than "finding structure in data" or "finding correlation in data". Those statistic terms just add another layer of thinking. Simply and creatively ask yourself practical questions, and then think about how to use data and basic analytical tasks to answer the questions. Of course we will definitely use visualization systems to assist the analytic tasks.

Snap-Together Visualization: Can Users Construct and Operate Coordinated Views?

North and Shneiderman, 2000

User study

Snap-Together Visualization: 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
    1. Detail-only
    2. Uncoordinated overview+detail
    3. Coordinated overview+detail

    Effects of adding overview to detail

    Effects of adding coordination
  - Task: 9 levels
    - A variety of browsing tasks, not grouped prior to the study
    - Tasks were closed-ended, with obvious correct answers
      e.g., "Which state has the highest college degree %"
      compare with "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-Together Visualization: user study

- Design
  - Within-subject, or everybody worked on all the interfaces/task combinations
  - Counterbalanced between interface (6 permutations) to avoid ordering / learning effect
    - In other words, had 3 main groups x 6 permutations = 18 participants
  - Need one task set (9) for each interface. The task in each set should be "identical"
  - Used the same task set order to avoid same grouping of interface and task set
  - Used the same task order within the set? (usually randomized)
  - 3 interfaces x 9 tasks = 27 tasks per study per participant
- Measurements
  - task completion time to obtain answer (no errors)
  - subjective ratings using rated scale (1-9)
- Participants:
  - 18 students, novice

Snap-Together Visualization: user study

- Time Result analysis: Hypothesis testing with ANOVA
  - 3 (interface) x 9 (task) within-subjects ANOVA to see if there were any main effects in terms of interface, or task, or interface/task interaction
    - ANOVA (ANalysis Of VAriance between groups)
      - A commonly used statistics for factorial designs
      - Tests the difference between the means of two or more groups, e.g., using a two-way ANOVA here to see if there is an effect on interface and task, or interaction
  - "Nine one-way ANOVAs reveal that user interface is significant for all nine tasks at $p < 0.0001$"
    - Not sure why they did this
  - "Individual t-tests between each pair of user interfaces within each task determine performance advantages"
    - This is post-hoc analysis, since ANOVA doesn't tell you which subpopulations are different
    - Need to correct for false positives due to multiple comparisons

Snap-Together Visualization: user study

- Time Result Analysis: Descriptive Statistics
  - E.g., "On average, coordination achieves an 80% speedup over detail-only for all tasks"
  - Good for discoveries based on results, e.g., the 3 task groups, and explain quantitative data with observed participant behaviours

- Subjective satisfaction analysis: Hypothesis testing with ANOVA
  - 3 (interface) x 4 (question category) within-subjects ANOVA
    - Usually do not use ANOVA for satisfaction score, as distribution may not be normal

Your Questions: trivial hypothesis? Valid comparisons?

- For one, the comparison of the baseline visualization of details-only is almost guaranteed to be trivially weaker than their coordinated interface as shown by many studies, and doesn't really serve to reinforce anything particularly new. Although this paper provides with a methodology of evaluating a new visualization technique (interaction), I want to raise a question that: do we always need to do empirical evaluation? Examining the second study, the result is so obvious and predictable. The experimented scope is simply about the amount users need to scroll. The second study has nothing to do with evaluating the performance of "Snap (or visualization coordination)" from many aspects or as a whole. It simply focuses on "overview and detail view coordination", which is only about reducing scrolling. Since there is no maneuver room of using single view, multiple views without coordination, and overview and detail coordination (if testers are instructed how to use each one of them effectively) in terms of scrolling for search, all the results of the task are predictable.

- The second study seems to merely compare oranges and apples to reinforce the strength of their system. While I do not doubt that their system has many benefits, I feel that the study presents the benefits in a forced manner that is not necessarily completely sound.

- Is it possible to classify papers on the type of evaluation/visualization? One class would be building an entirely new interface, i.e. this paper? Another class would be a very focused and controlled subject where it attempts to mold a theory such as fitts law. This paper would be classified as building and evaluating an entirely new interface because it uses several theories to build a utility called snap.

Your Questions: statistical analysis

- It appears that they run nine one-way ANOVA's and multiple t-tests between pairs of interfaces and state that performance advantages were revealed by these individual tests. If I am not mistaken, doing this many individual tests is bad practice as it significantly raises the probability that at least one of them is a false positive.

- How important are statistical significance tests in user studies? Snap-together makes use of them, but a lot of papers don't. Is it common to test the significance of results and not report the significance level, or are significance tests often not done?

Your Questions: user behaviours

- I find it very strange that the authors were surprised that users found scrolling through a large textual report to be difficult - I'm tempted to ask whether they (the authors) have ever used a web browser! In my experience, at least, the task of finding a particular piece of information solely by scrolling through a long web page is not only cognitively difficult but also has a low success rate, for both novice and advanced users. In fact, it doesn't strike me as being significantly easier than searching through a book (with which one is not very familiar).

In Summary: Two evaluation techniques

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