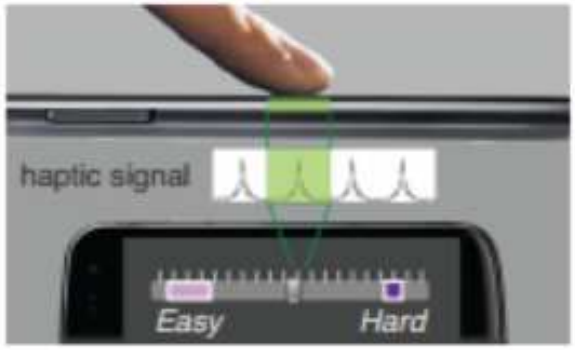
Empirical Methods

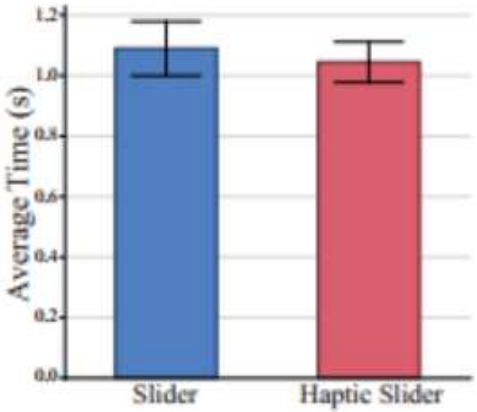
Parallel Universes

- A note on alt.CHI papers …
- Simulated running an experiment in multiple universes
  - Note: Really just ran the experiment eight times
  - Note: Actually just simulated the experiment eight times based on generic distribution of results drawn from a sample (see discussion).

Experimental Design

- A repeated measure full-factorial within-subject design was used.
- The factors were Technique = S=slider, HS=haptic slider, and Difficulty = Easy, Hard.
- Twelve volunteers (2 female) familiar with touch devices, aged 22-36, participated in the study. We collected a total of 12 Participant X 2 Technique X 2 Difficulty X 128 repetitions = 6144 trials with completion Time.





Comments

- I like the idea of running studies in parallel universes which would give a better view of how people behave; even though in this paper, it seems to me they are just doing replication studies with different groups of people. (Edwin)
- No solution to this dilemna is suggested or ‘the experiment should have ran in 9 parallel universes so it could uncover more problems.’ (Jeff, Hemant, Valerie, Shaishav)
- Connor: The treatment of the arbitrary cutoff of 0.05 may need to be reconsidered.

Modeling Human Performance of Pen Stroke Gestures

- Context: Shuman Zhai invented shapewriter.
  - Previously know as SHARK, Shorthand-Aided Rapid Keyboarding
  - Swype is a variant
- Wants to model gestures
  - Expert level performance
  - Enhanced recognition
  - Etc.
- Proposes a CLC model for characters

$$T = \sum T(line) + \sum T(corner) + \sum T(curve)$$

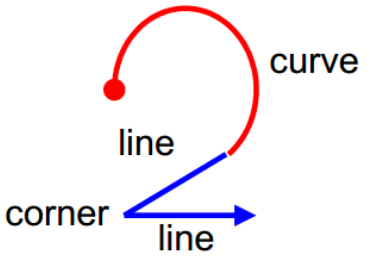


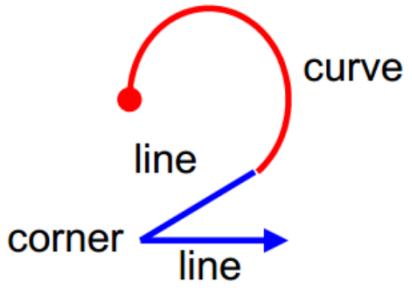
Figure 3. Decomposition of a gesture.

What did Cao and Zhai do?

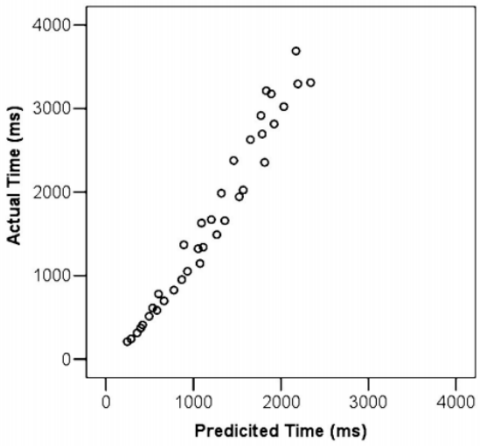
- Leveraged one model of movement, 2/3 power law, for curved strokes
  - Called it the “power law” and did not use 2/3 coefficient …
- Derived model for straight lines using another power law
- Analyzed corners to test time
- Found:
  - $T(\text{line}) = 68.8\ L^{0.469}$
  - $T(\text{arc}) = \alpha\ r^{1-0.586} / 0.0153$
  - T(corner) => break the line into two components

Results

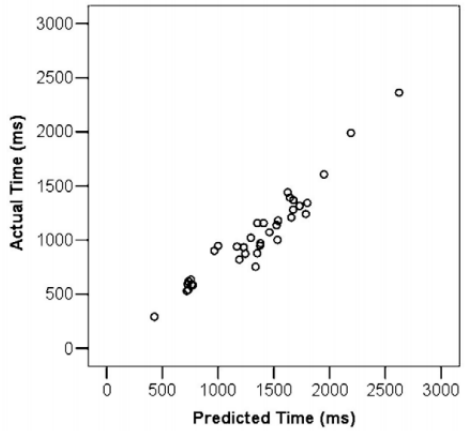
- Take a shape like the 2 on the right
- Make participants draw the shape within an accuracy constraint
- Found good agreement with model initially
  - Note, however, that polylines underestimate, and arbitrary lines overestimate



Polylines

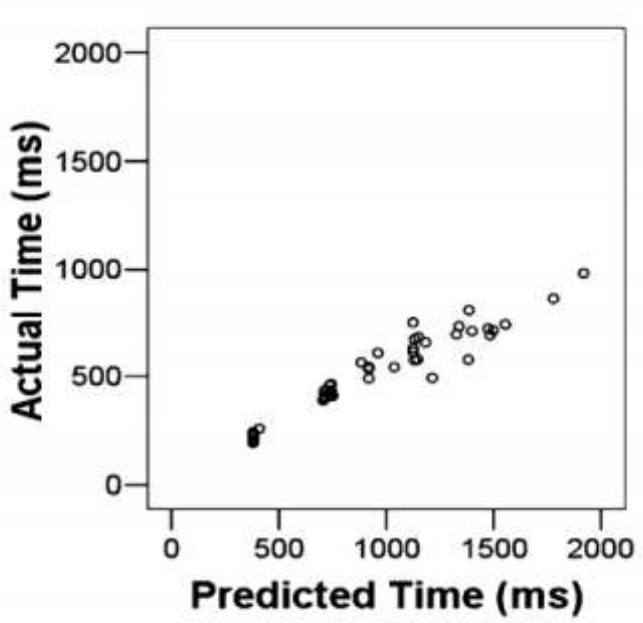


Arbitrary lines

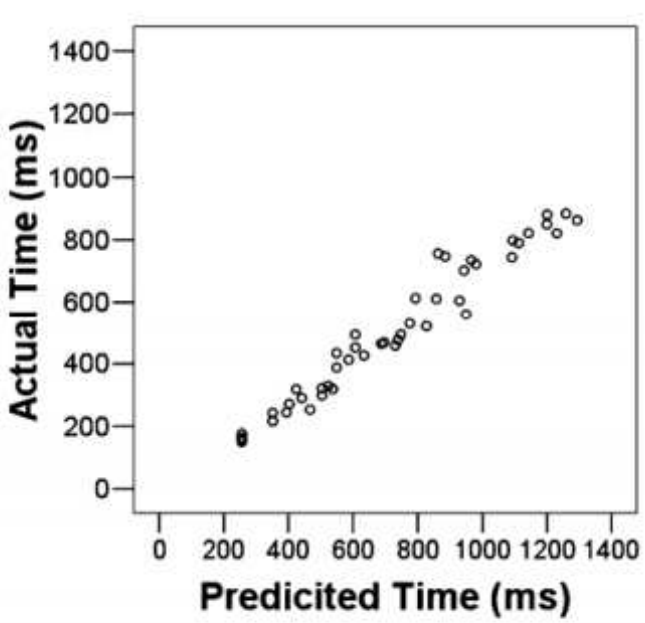


Testing: Unistrokes and Shapewriter

- Model generally over-predicted time, though correlation was good … maybe



(a) Unistrokes & Graffiti



(b) ShapeWriter gestures

Discussion

- Density of results section (Connor, Valerie, Jeff)
- Confounds:
  - Habits of using touchscreen devices for writing purposes (Shaishav)
  - Range of sizes small, different relationship between size and completion time if the gestures require more elbow and shoulder movement (Valerie) or variability in gesture (Edwin)
  - Mental complexity which could have been tested with the tools such as NASA – TLX (Hemant)

Discussion: Over-estimation of time

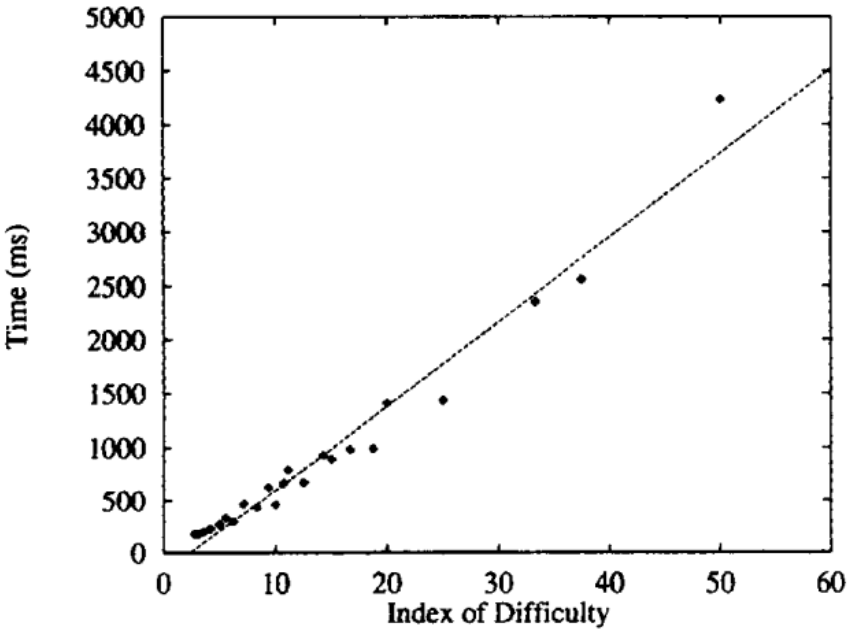


Figure 5: Scatter-plot of the MT-ID relationship. The relation fitted was $MT = a + b \times ID$ where $ID = \frac{A}{W}$

From Accot and Zhai

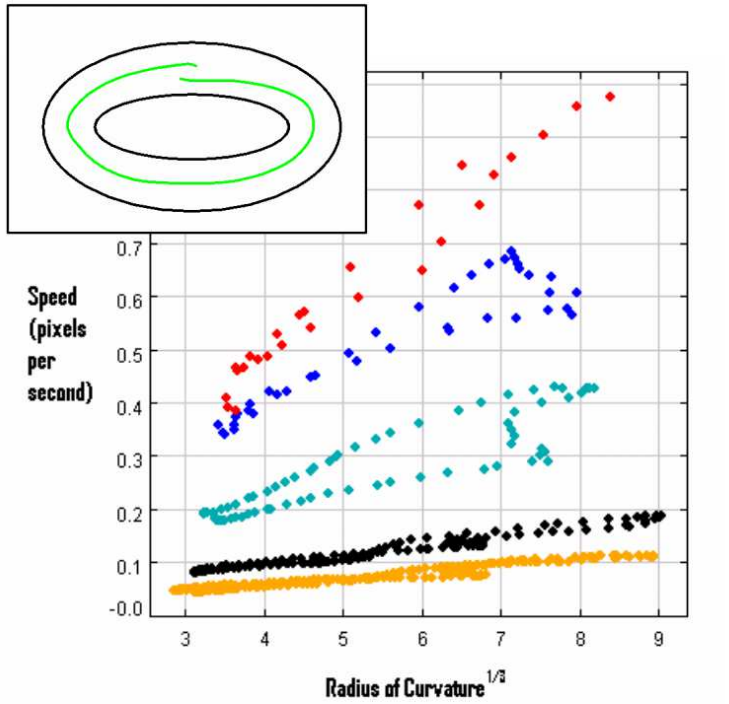


Figure 6: Speed vs. Radius of Curvature$^{1/3}$ under constraints of 80, 60, 40, 20, and 10 pixel tunnels for a single user.

From Lank and Saund citation

I really want someone to validate the V(s) α W(s) r(s)$^{1/3}$

Empirical Methods

$t = a + b$

Latin Square Design

A	*B*	*C*	*D*	*E*
B	*A*	*D*	*E*	*C*
C	*E*	*A*	*B*	*D*
D	*C*	*E*	*A*	*B*
E	*D*	*B*	*C*	*A*

Overview: Empirical Methods

- Wikipedia
  - Any research which bases its findings on observations as a test of reality
  - Accumulation of evidence results from planned research design
  - Academic rigor determines legitimacy
- Frequently refers to scientific-style experimentation
  - Many qualitative researchers also use this term

Positivism

- Describe only what we can measure/observe
  - No ability to have knowledge beyond that
- Example: psychology
  - Concentrate only on factors that influence behaviour
  - Do not consider what a person is thinking
- Assumption is that things are deterministic

Post-Positivism

- A recognition that the scientific method can only answer question in a certain way
- Often called critical realism
  - There exists objective reality, but we are limited in our ability to study it
  - I am often influenced by my physics background when I talk about this
    - Observation => disturbance

Implications of Post-Positivism

- The idea that all theory is fallible and subject to revision
  - The goal of a scientist should be to disprove something they believe
- The idea of triangulation
  - Different measures and observations tell you different things, and you need to look across these measures to see what's really going on
- The idea that biases can creep into any observation that you make, either on your end or on the subject's end

Experimental Biases in the RW

- Hawthorne effect/John Henry effect
- Experimenter effect/Observer-expectancy effect
- Pygmalion effect
- Placebo effect
- Novelty effect

Hawthorne Effect

- Named after the Hawthorne Works factory in Chicago
- Original experiment asked whether lighting changes would improve productivity
  - Found that anything they did improved productivity, even changing the variable back to the original level.
  - Benefits stopped or studying stopped, the productivity increase went away
- Why?
  - Motivational effect of interest being shown in them
- Also, the flip side, the John Henry effect
  - Realization that you are in control group makes you work harder

Experimenter Effect

- A researcher's bias influences what they see
- Example from Wikipedia: music backmasking
  - Once the subliminal lyrics are pointed out, they become obvious
- Dowsing
  - Not more likely than chance
- The issue:
  - If you expect to see something, maybe something in that expectation leads you to see it
- Solved via double-blind studies

Pygmalion effect

- Self-fulfilling prophecy
- If you place greater expectation on people, then they tend to perform better
- Studied teachers and found that they can double the amount of student progress in a year if they believe students are capable
- If you think someone will excel at a task, then they may, because of your expectation

Placebo Effect

- Subject expectancy
  - If you think the treatment, condition, etc has some benefit, then it may
- Placebo-based anti-depressants, muscle relaxants, etc.
- In computing, an improved GUI, a better device, etc.
  - Steve Jobs: http://www.youtube.com/watch?v=8JZBLjxPBUU
  - Bill Buxton: http://www.youtube.com/watch?v=Arrus9CxUiA

Novelty Effect

- Typically with technology
- Performance improves when technology is instituted because people have increased interest in new technology
- Examples: Computer-Assisted instruction in secondary schools, computers in the classroom in general, smartwatches (particularly the Apple Watch).

What can you test?

- Three things?
  - Comparisons
  - Models
  - Exploratory analysis
- Reading was comparative with some nod to model validation

Concepts

- Randomization and control within an experiment
  - Random assignment of cases to comparison groups
  - Control of the implementation of a manipulated treatment variable
  - Measurement of the outcome with relevant, reliable instruments
- Internal validity
  - Did the experimental treatments make the difference in this case?
- Threats to validity
  - History threats (uncontrolled, extraneous events)
  - Instrumentation threats (failure to randomize interviewers/raters across comparison groups
  - Selection threat (when groups are self-selected)

Themes

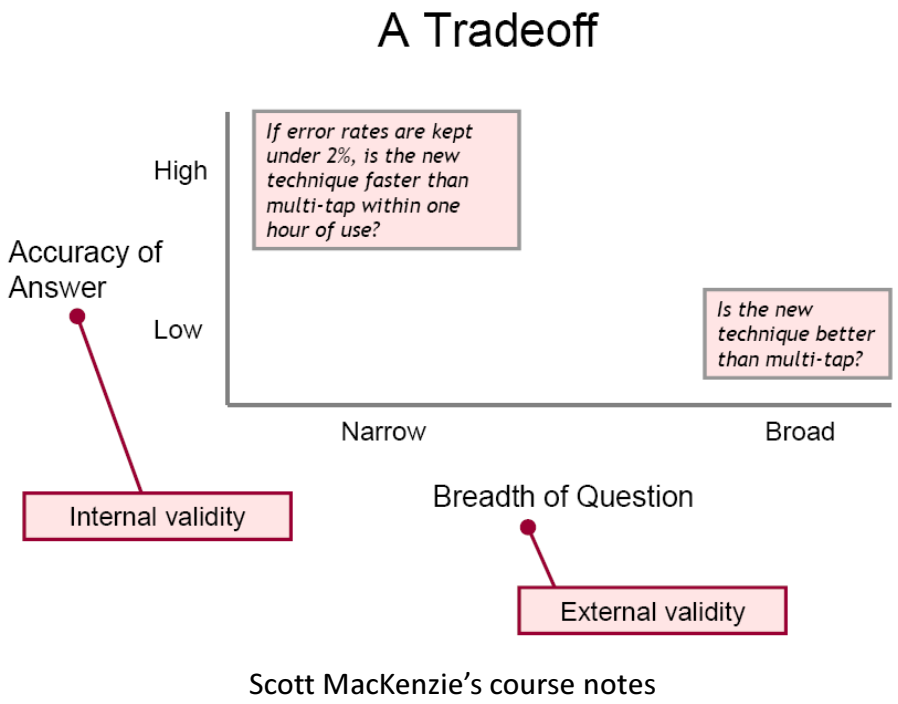
- HCI context
- Scott MacKenzie's tutorial
  - Observe and measure
  - Research questions
  - User studies – group participation
  - User studies – terminology
  - User studies – step by step summary
  - Parts of a research paper

Observations and Measures

- Observations
  - Manual (human observer)
    - Using log sheets, notebooks, questionnaires, etc.
  - Automatically
    - Sensors, software, etc.
- Measurements (numerical)
  - Nominal: Arbitrary assignment of value (1=male, 2=female
  - Ordinal: Rank (e.g. 1^{st}, 2^{nd}, 3^{rd}, etc.
  - Interval: Equal distance between values, but no absolute zero
  - Ratio: Absolute zero, so ratios are meaningful (e.g. 40 wpm is twice as fast as 20 wpm typing)
- Given measurements and observations, we:
  - Describe, compare, infer, relate, predict

Research Questions

- You have something to test (a new technique)
- Untestable questions:
  - Is the technique any good?
  - What are the technique's strengths and weaknesses?
  - Performance limits?
  - How much practice is needed to learn?
- Testable questions seem narrower
  - See example at right



Scott MacKenzie's course notes

Research Questions (2)

- Internal validity
  - Differences (in means) should be a result of experimental factors (e.g. what we are testing)
  - Variances in means result from differences in participants
  - Other variances are controlled or exist randomly
- External validity
  - Extent to which results can be generalized to broader context
  - Participants in your study are “representative”
  - Test conditions can be generalized to real world
- These two can work against each other
  - Problems with “Usable”

Research Questions (3)

- Given a testable question (e.g. a new technique is faster) and an experimental design with appropriate internal and external validity
- You collect data (measurements and observations)
- Questions:
  - Is there a difference
  - Is the difference large or small
  - **Is the difference statistically significant**
  - Does the difference matter

Significance Testing

- R. A. Fisher (1890-1962)
  - Considered designer of modern statistical testing
- Fisher's writings on Decision Theory versus Statistical Inference:
  - An important difference is that Decisions are final while the state of opinion derived from a test of significance is provisional, and capable, not only of confirmation but also of revision (p.100).
  - A test of significance ... is intended to aid the process of learning by observational experience. In what it has to teach each case is unique, though we may judge that our information needs supplementing by further observations of the same, or of a different kind (pp. 100-101).
- Implications?
  - What is the difference between statistical testing and qualitative research?

Testing

- Various tests
  - t- and z-tests for two groups
  - ANOVA and variants for multiple groups
  - Regression analysis for modeling
- Also
  - Binomial test for distributions
  - CHI-Square test for tabular values
- Great on-line resources:
  - http://www.statisticshell.com/
  - http://www.statisticshell.com/html/limbo.html
  - Jacob Wobbrock's tutorial

Research Design

- Participants
  - Formerly “subjects”
  - Use appropriate number (e.g. similar to what others have used)
- Independent variable
  - What you manipulate, and what levels of iv were tested (test conditions)
- Confounding variables
  - Variables that can cause variation
  - Practice, prior knowledge

Research Design (2)

- Within subjects versus between subjects
  - Within = repeated measures
  - Sometimes a choice:
    - Controls subject variances (easier stat significance), but can have interference
- Counterbalancing
  - Typing on qwerty versus numeric keyboard
    - Could learn phrases, some phrases could be easier, so vary order of devices
  - Latin square
  - http://www.yorku.ca/mack/RN-Counterbalancing.html

Reading Experimental Results

- Sometimes you need to read carefully to fully appreciate what data is saying
- Worked example: Wedge

Wedge

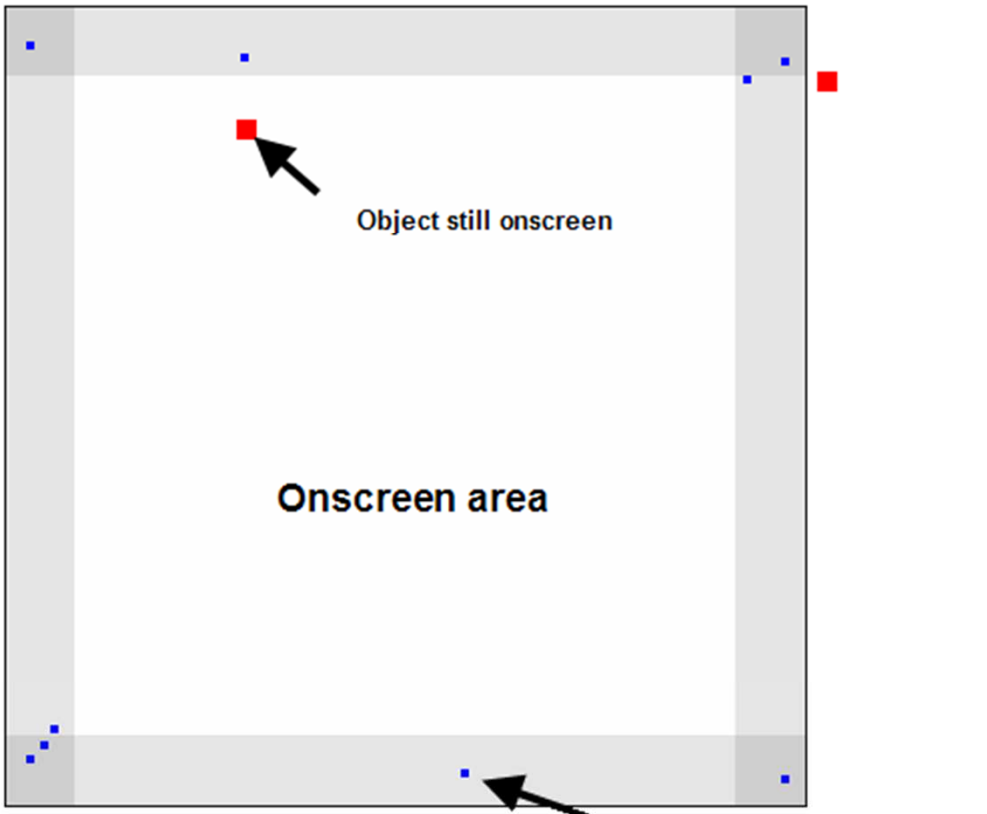
http://patrickbaudisch.com/projects/wedge/

To overcome display limitations of small-screen devices, researchers have proposed techniques that point users to objects located off-screen. Arrow-based techniques such as City Lights convey only direction. Halo conveys direction and distance, but is susceptible to clutter resulting from overlapping halos. We present Wedge, a visualization technique that conveys direction and distance, yet avoids overlap and clutter. Wedge represents each off-screen location using an acute isosceles triangle: the tip coincides with the off-screen locations, and the two corners are located on- screen. A wedge conveys location awareness primarily by means of its two legs pointing towards the target. Wedges avoid overlap programmatically by repelling each other, causing them to rotate until overlap is resolved. As a result, wedges can be applied to numbers and configurations of targets that would lead to clutter if visualized using halos. We report on a user study comparing Wedge and Halo for three off-screen tasks. Participants were significantly more accurate when using Wedge than when using Halo.

Related Work

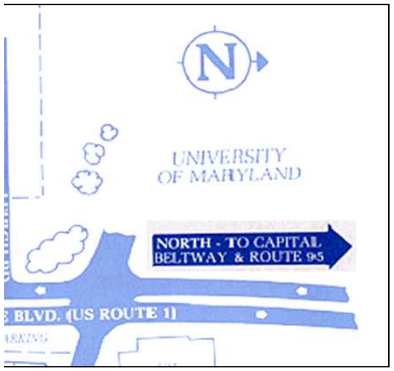
- Edgeradar
- Arrows
- City lights
- Halo

edgeradar



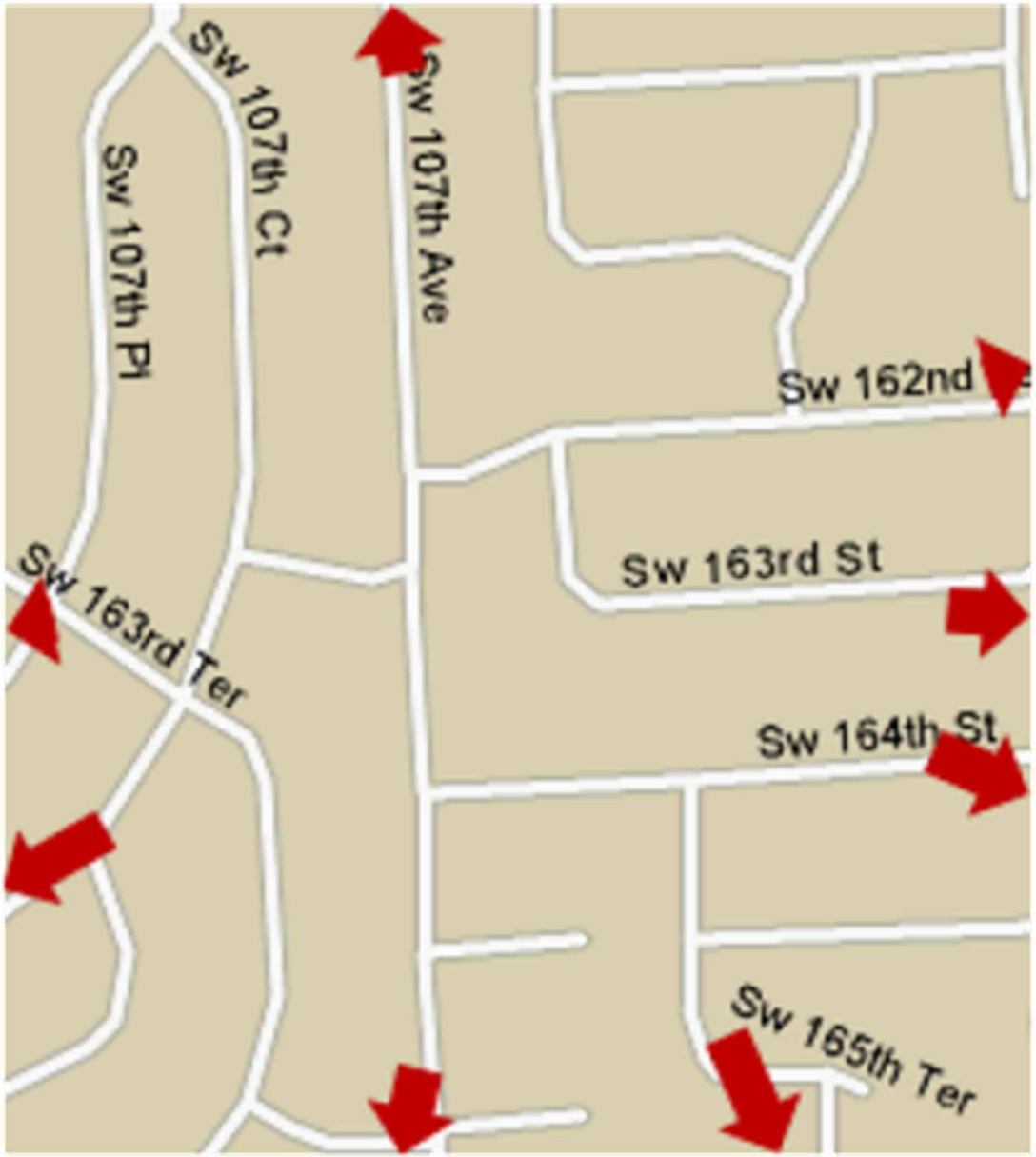
[Gustafson 07]

simple arrows



[Tecmo Bowl 87]

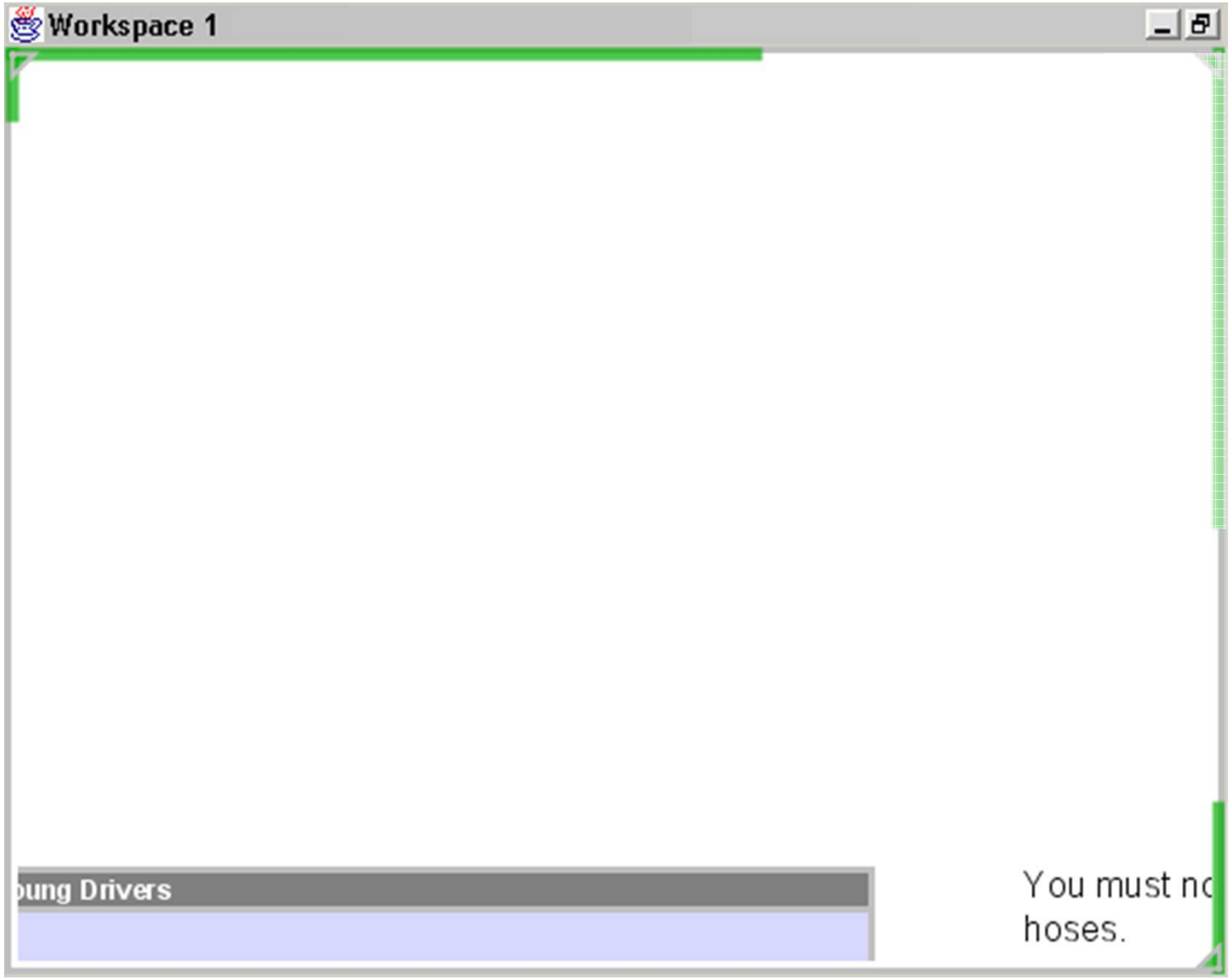
scaled and stretched arrows



[Burigat 06]

city lights

"space-efficient fisheye technique"



[Mackinlay 03]

halo

[Baudisch 03]

Related Work

- Edgeradar
- Arrows
- City lights
- Halo
- Problem with halo:
  - Clutter and corners


COMPAQ
iPAQ
pocket pc
Virginia St
Walnut St
Shattuck Ave
Grant St
Delaware St
University Ave
Berkeley
Addison St
Martin Luther
King Jr
Memorial Park
McGee Ave
Jefferson Ave
Roosevelt Ave
Bancroft Way
Harold Way
Milvia St
Shattuck Ave
Durant Ave
Haste St

Evaluation

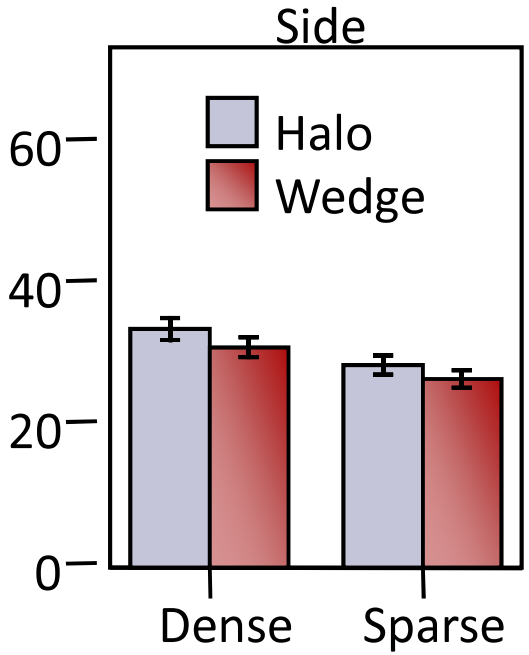
- 18 subjects, with 2 removed because of high error rate
  - Note: This is OK …
- Three tasks:
  - Locate: Click off-screen where you think the target is
  - Avoid: Traffic jams are indicated and you need to click the hospital furthest from traffic jams
  - Closest: Click on halo/wedge corresponding to closest off-screen location

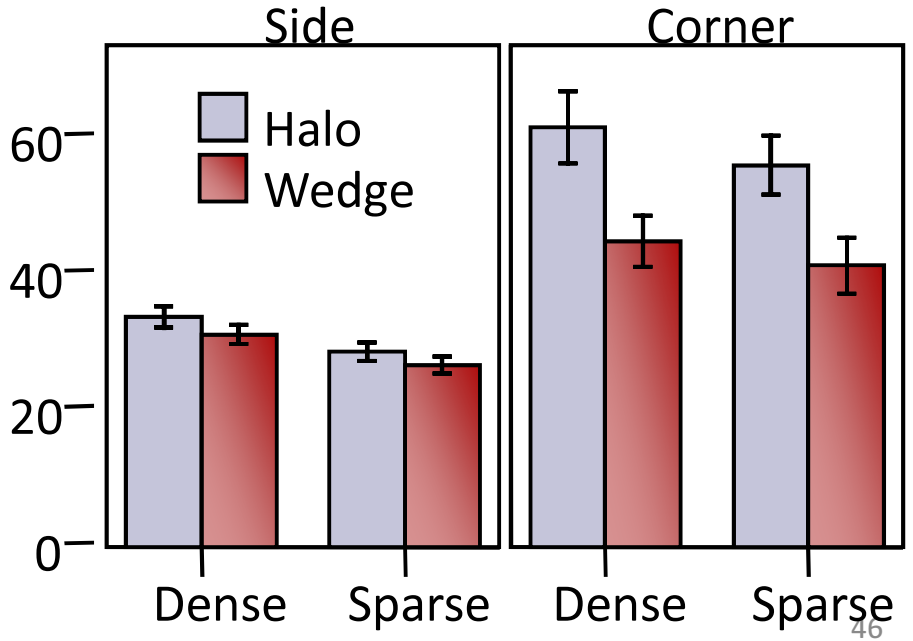
Hypotheses

- **Wedge is more accurate**
- Larger improvement in dense condition
- **Larger improvement in corners**
  - (no hypothesis about task time)

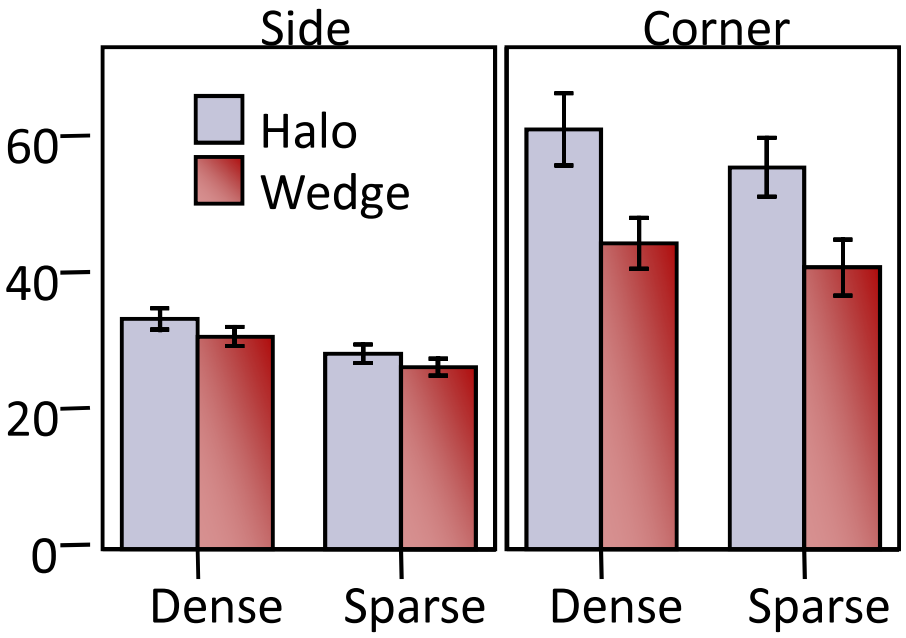
Results

- No significant difference in task time
- Participants were significantly more accurate when using the wedge





Locate Task



As can be seen from Figure 11 larger errors were seen in corner trials (mean 51 pixels) than in side trials (mean 30 pixels). There were also larger errors in dense configurations (mean 43) than sparse configurations (mean 38). The overall difference between visualizations was about 10 pixels (Halo mean 45.3 pixels; Wedge mean 35.6 pixels).

In addition, there was a significant interaction between Visualization and Position ($F_{1,15}=15.36$, $p=0.001$). As shown in Figure 11, the difference between visualization types is considerably larger in corners than on the sides of the screen, which supports our hypothesis that the reduced space in corners causes additional problems for Halo interpretation. There was no interaction between Visualization and Density ($F_{1,15}=0.67$, $p=0.43$).

Additional Results

Avoid:

Figure 13 shows error rates for the different visualizations, densities, and positions. A 2x2x2 ANOVA did not show any effects of Visualization ($F_{1,15}=2.55$, $p=0.13$), Position ($F_{1,15}=2.38$, $p=0.14$), or Density ($F_{1,15}=0.58$, $p=0.46$). In addition, there were no interactions between any factors.

A 2x2x2 ANOVA showed no effects of any of the three factors on task completion time (Visualization $F_{1,15}=0.18$, $p=0.68$; Density $F_{1,15}=2.09$, $p=0.17$; Position $F_{1,15}=1.58$, $p=0.23$), and no interactions between any factors.

Closest

Figure 15 shows error rates for the different visualizations, densities, and positions. A 2x2x2 ANOVA showed significant main effects of Position ($F_{1,15}=76.6$, $p<0.001$)), but not of Visualization ($F_{1,15}=1.24$, $p=0.28$) or Density ($F_{1,15}=0.12$, $p=0.73$). There was a significant interaction between Density and Position ($F_{1,15}=7.33$, $p=0.016$), but no interactions with Visualization.

A 2x2x2 ANOVA showed significant main effects of Position ($F_{1,15}=5.24$, $p=0.037$), but did not show effects of Visualization ($F_{1,15}=0.10$, $p=0.76$) or Density ($F_{1,15}=2.89$, $p=0.11$). There was, however, a significant interaction between Visualization and Density ($F_{1,15}=6.60$, $p=0.021$).

Additional Results

	Wedge	Halo	No Preference
Locate	10	5	1
Avoid	10	5	1
Closest	6	8	2

Table 3: The number of participants who preferred each visualization technique for the three tasks.

Comments made during the trial suggested reasons for the advantages for Wedge over Halo. One user said, “I found that when the rings overlap it is almost impossible to tell which is the right ring. Wedges just seem natural.” And another stated, “overlapping rings made it very confusing at times. Directional wedges helped a lot, and they also seem to take up less space. More information meant less thinking with the wedges.” Participant’s comments also provided some insight into the reasons why Halo was preferred for the Closest task – that the difference between distant and close off-screen objects was easier to determine with Halo, since there is a large visual difference in this case. One participant stated that, “the sizes of the arcs did not require too much calculation or thinking to spot the smallest ring.”

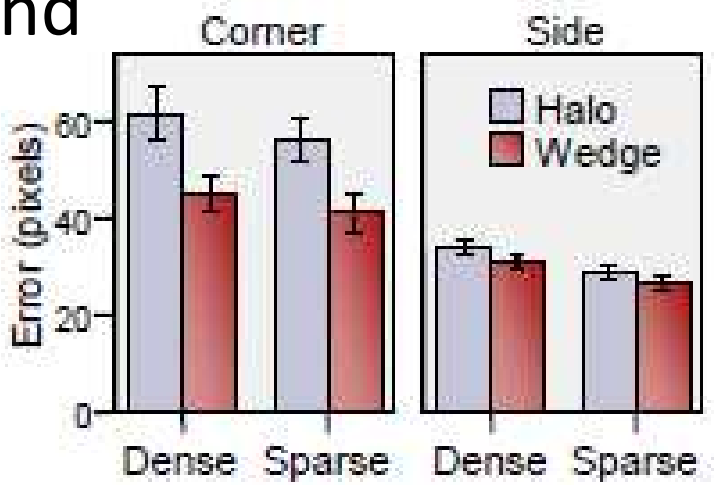
Meta-Level Comments: Experimental Papers

- A lot of techniques + evaluation
- Predictable outline:
  - Problems with existing techniques
  - Rationale for new design
  - Evaluation of new design
    - Usually two or three tasks
  - Discussion and implications

Your thoughts?

My Problem with Wedge

- Read the paper
- For visualization, ONLY LOCATE had significant differences, and ONLY FOR ERROR
- But 2 participants were removed for high error …
- And note that, IMO, visualization is only significant for corners



Second consideration

- Closest completion time was the only other area of significance, and only for interactions
- A 2x2x2 ANOVA showed significant main effects of Position ($F_{1,15}=5.24$, $p=0.037$), but did not show effects of Visualization ($F_{1,15}=0.10$, $p=0.76$) or Density ($F_{1,15}=2.89$, $p=0.11$). There was, however, a significant interaction between Visualization and Density ($F_{1,15}=6.60$, $p=0.021$).

- Problem:
  - Why not explore this interaction as they do for errors in Locate?

Concerning because

Figure 16. Closest task mean completion time. Error bars indicate standard error.

Another problem

- Graphs
  - Kept on showing dense-sparse for Halo-Wedge even when no interactions
  - Particular problem in locate because of interaction between density and position, but not visualization:

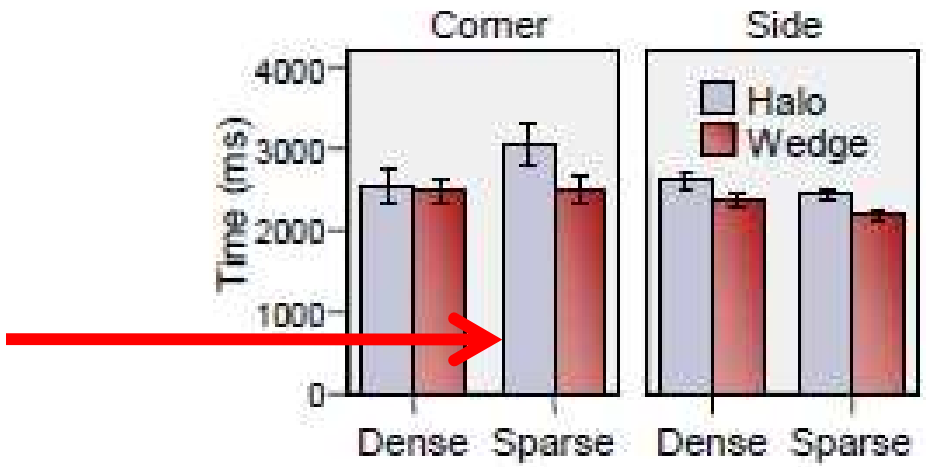


Figure 12. Locate task mean completion time. Error bars indicate standard error.