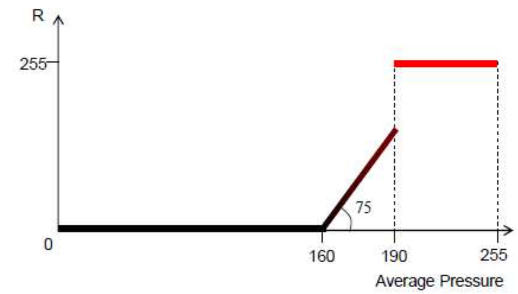
Empirical Methods

Experimental Analysis of Mode Switching Techniques, Yang Li et al.

- Premise: Mode switching is an important and common task in pen-tablet interfaces
  - Based on the need to overload pen behavior
  - So evaluate different techniques to see which one is better in speed and errors
- Five mode switching techniques
  - Button in toolbar, press-and-hold, non-preferred hand, pressure, flipping pen

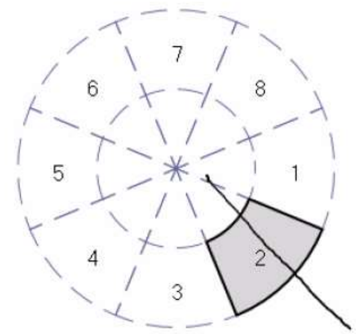
Methodology

- First a pilot
  - Tests pressure levels
- Experimental task
  - Pie cutting task
    - Baseline and compound (control and experimental)



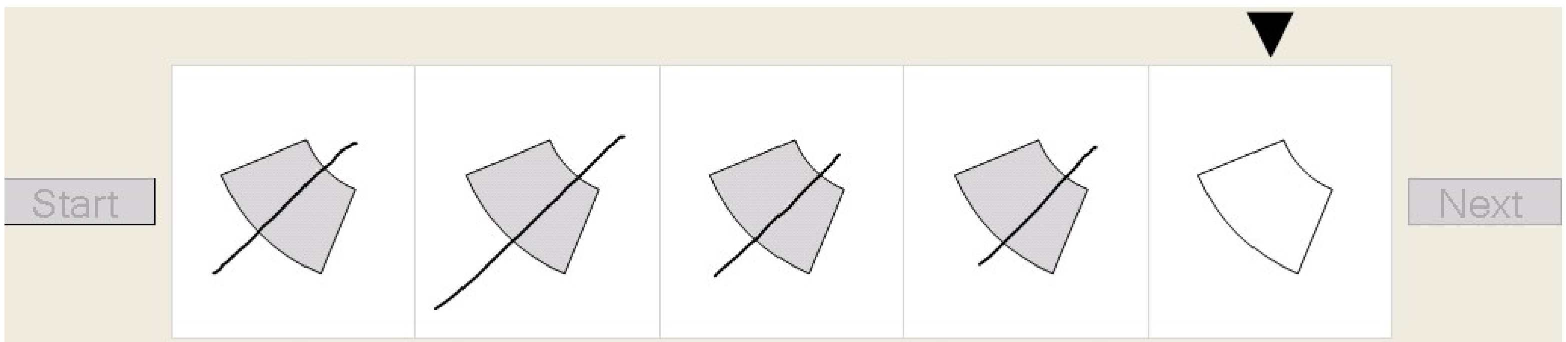
To capture the nature of sketch-based interactions, which are normally informal and fluid, we designed a pie-crossing task as an abstraction of the action of gesturing and inking (see Figure 2). A pie slice is shown with one of eight orientations corresponding to the eight major geographical directions. A participant was required to quickly cross a slice from its inner edge towards its outer edge according to a target's orientation. This design examines the drawing of various directions without requiring precise positioning and careful alignment by participants. This design also captures a realistic use scenario of gestures, i.e., marking menus [8], where users can cross a series of objects with marks to perform different commands.

Task

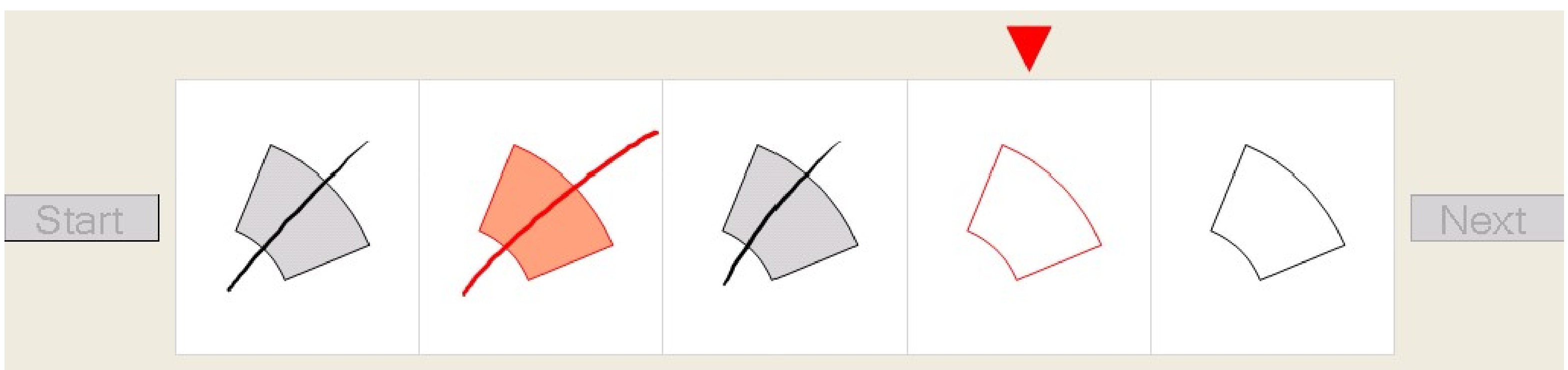


- Pie cutting

Control



Experimental



Errors

- One set of measures is accuracy of mode switching
- What are errors?
- Mode errors
  - Mode-in and Mode-out
- Crossing errors
- Out of target errors

Procedure

- Training phase + test phase
  - In this case training data ignored
- 5X5 Latin square counter-balanced techniques
- 9 Blocks
  - First baseline/control, then experimental, alternating for 5 control, 4 experimental

The experiment included a training phase for the baseline tasks, five experimental sessions with one session for each technique, and a post-study questionnaire. The experiment took about 80 minutes in total. A 5x5 Latin Square was used to counterbalance the order of the techniques. Each session was divided into two parts. The first part involved learning to use a mode switching technique and extensive practice. The second part was the experimental phase in which a participant was given 9 blocks of trials. The first block was a *baseline task* and then a *compound task*, alternating until the ninth block ended with a *baseline task*. A participant could take a break between blocks. In total, the experiment consisted of:

15 participants x
5 mode switching techniques x
9 block of trials x
8 screens (8 orientations) x
5 pie-crossing tasks
= 27,000 pie-crossing tasks

Counter-balanced: 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*

Measures

- Dependent variables are time, errors, preference
- In 9 experimental blocks
  - Two used as warm up, seven analysed
- Duration divided into 3 cycles
  - Break after first, third, last pie
  - Last two cycles have a mode-switch
- Mode switch time = average cycle duration for last two cycles with mode-switch in compound – average cycle duration for last two cycles in control (see slide 4)

The dependent variables were the *mode switching time*, the total number of errors in a compound task, and the subjective preference of participants. The first two blocks in the experimental phase were for warming up and the data of the seven following blocks were used for analysis.
The timing for each screen is started when the Start button is clicked and automatically ended when the last pie is crossed and the pen is lifted. This duration is divided into three cycles. The first cycle starts when the Start button is clicked and ends when the first pie is crossed. The second cycle starts right after the first cycle and ends after the third pie is crossed. This is followed by the third cycle, which includes crossing the last two slices. Therefore, one target needs to be crossed in the first cycle and two targets need to be crossed in each of the second and the third cycles. We call cycle 2 and 3 *full cycles* and cycle one the *start cycle*.
In a *compound task*, a full cycle contains a complete mode switch process including switching into gesture mode and switching back to ink mode.
The *mode switching time* for each of the three compound blocks was computed by subtracting the mean of the two adjacent baseline tasks' *average cycle durations* from the compound block's *average cycle duration*. Average cycle duration was the mean duration of all correct full cycles in a block.

Results

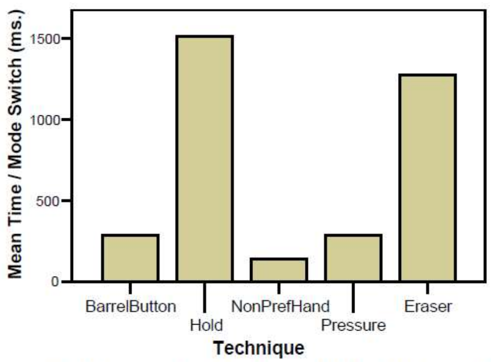


Figure 5. The mean mode switching time of five techniques.

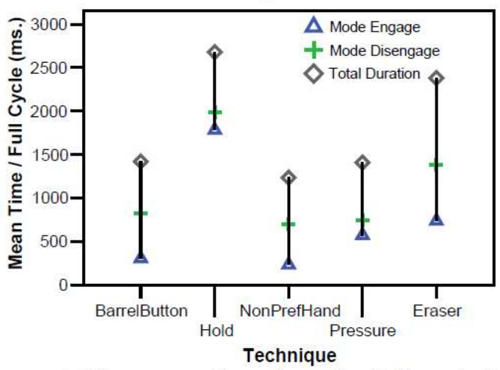


Figure 6. The mean duration of a full cycle in a compound task (before subtraction). The time of mode engagement and disengagement are also shown.

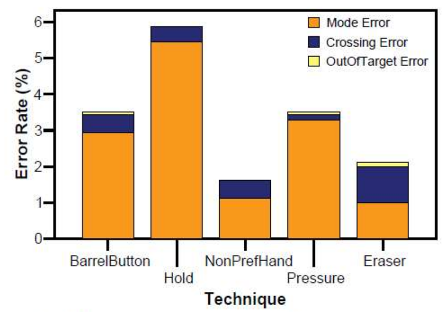


Figure 7. The error rate on each pie-crossing. The mode errors included both Mode-In and Mode-Out errors.

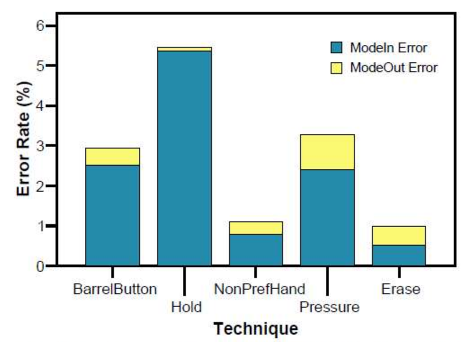


Figure 8. The occurrence rate of Mode-In and Mode-Out errors on each pie-crossing, which is weighted by the frequency of inking and gesturing, respectively. 40% of pie-crossings were for gesturing, while 60% were for inking.

Table 1. The participants' preferences for each technique.

Dimension	Barrel-Button	Hold	NonPref-Hand	Pressure	Eraser
Learning	4.4	3.5	4.7	3.5	4.2
Use	3.7	2.2	4.1	3.4	2.4
Accuracy	3.7	2.9	4.6	3.3	3.6
Speed	4	1.7	4.5	4	1.9
Eye fatigue	4.1	3.3	4.4	3.9	4.2
Hand fatigue	3.5	3.3	4.1	3.3	2.1

Empirical Methods
$t = a + b$

Research Landscape

- Quantitative = Positivist/post-positivist approach
  - Evaluate hypotheses via experimentation
- Qualitative = Constructivist approach
  - Build theory from data

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, etc.

What can you test?

- Three things?
  - Comparisons
  - Models
  - Exploratory analysis
- Reading was comparative

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

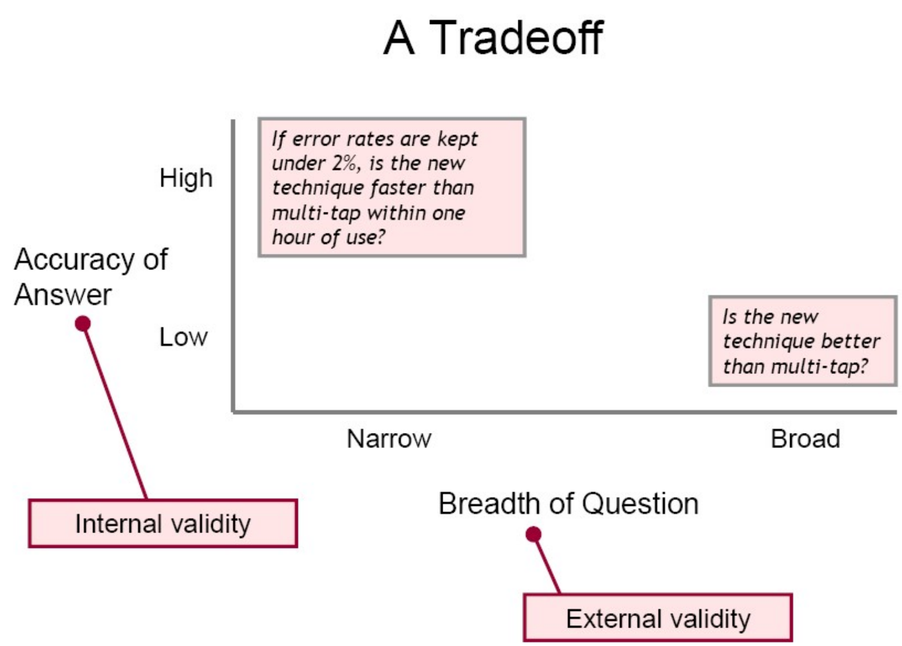
- HCI context
- Scott MacKenzie’s tutorial
  - Observe and measure
  - Research questions
  - User studies – group participation
  - User studies – terminology
  - User studies – step by step summary
  - Parts of a research paper

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"
  - Noted by many with the readings

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