Empirical Research Methods in Human-Computer Interaction

Scott MacKenzie
York University

Part I – The Short Answer

What is Empirical Research?

- Empirical research is…
  - observation-based investigation seeking to discover and interpret facts, theories, or laws.

Why do Empirical Research?

- Two answers…
- First, we conduct empirical research to…
  - answer (or raise!) questions about a new or existing UI design or interaction method.
- Second… (we'll get to this later)

How do we do Empirical Research?

- We conduct empirical research through…
  - a program of inquiry conforming to "the scientific method".

Part II – The Long Answer (with an HCI context)

Three Themes

- Answer and raise questions
- Observe and measure
- User studies

Observe and Measure

- Observations are gathered…
  - Manually (human observers)
  - Automatically (computers, software, sensors, etc.)
- A measurement is a recorded observation

When you cannot measure, your knowledge is of a meager and unsatisfactory kind.

Kelvin, 1883

Scales of Measurement

- Nominal
- Ordinal
- Interval
- Ratio

crude ↕ sophisticated

<u>Nominal</u> – arbitrary assignment of a code to an attribute, e.g.,
1 = male, 2 = female

<u>Ordinal</u> – rank, e.g.,
1st, 2nd, 3rd, …

<u>Interval</u> – equal distance between units, but no absolute zero point, e.g.,
20° C, 30° C, 40° C, …

<u>Ratio</u> – absolute zero point, therefore ratios are meaningful, e.g.,
20 wpm, 40 wpm, 60 wpm

Use ratio measurements where possible

9

Ratio Measurements

- Preferred scale of measurement
- With ratio measurements summaries and comparisons are strengthened
- Report "counts" as ratios where possible
- Example – a 10-word phrase was entered in 30 seconds
  - Bad: t = 30 seconds
  - Good: Entry rate = 10 / 0.5 = 20 wpm
- Example – two errors were committed while entering a 10-word (50 character) phrase
  - Bad: n = 2 errors
  - Good: Error rate was 2 / 50 = 0.04 = 4%

10

Research Questions

- Why do we conduct empirical research?
- Simply…
  - To answer (or raise!) questions about a new or existing UI design or interaction technique!
- Questions include…
  - Is it viable?
  - Is it as good as or better than current practice?
  - Which of several design alternatives is best?
  - What are its performance limits and capabilities?
  - What are its strengths and weaknesses?
  - Does it work well for novices, for experts?
  - How much practice is required to become proficient?

11

Testable Research Questions

- Preceding questions, while unquestionably relevant, are not testable
- Try to re-cast as testable questions (…even though the new question may appear less important)

Scenario…

You have invented a new text entry technique for mobile phones. In your view, it's pretty good. In fact, you think it's better than the most widely used current technique, multi-tap. You decide to undertake some empirical research to evaluate your invention and to compare it with multi-tap? What are your research questions?

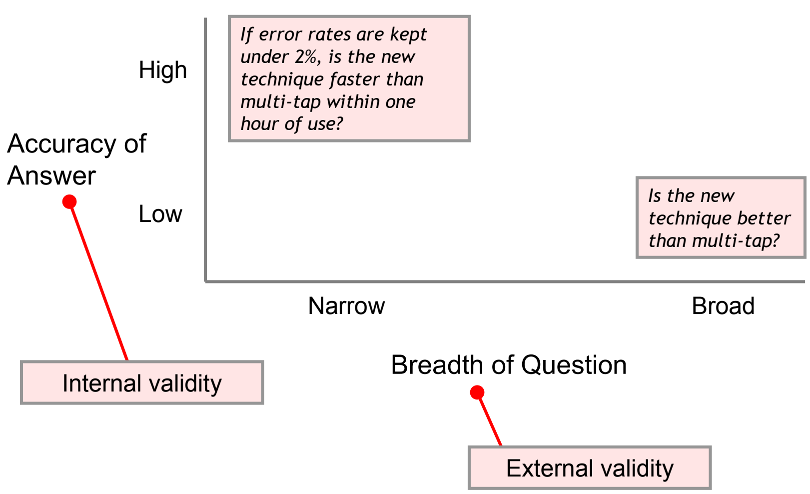
12

Research Questions (2)

- Weak question…
  - *Is the new technique better than multi-tap?*
- Better…
  - *Is the new technique faster than multi-tap?*
- Better still…
  - *Is the new technique faster than multi-tap within one hour of use?*
- Even better…
  - *If error rates are kept under 2%, is the new technique faster than multi-tap within one hour of use?*

York University – Department of Computer Science and Engineering

A Tradeoff



Internal Validity

- Definition: The extent to which the effects observed are due to the test conditions
- Statistically…
  - Differences in the means are due to inherent properties of the test conditions
  - Variances are due to participant differences ('pre-dispositions')
  - Other potential sources of variance are controlled or exist equally and randomly across the test conditions
  - Note: Uncontrolled sources of variance are potentially bad news and may compromise internal validity (see "confounding variable" later).

External Validity

- Definition: The extent to which results are generalizable to other people and other situations
- Statistically…
  - Re people, the participants are representative of the broader intended population of users
  - Re situations, ***Test environment*** and ***experimental procedures*** are representative of real world situations where the UI/technique will be used

Test Environment Example

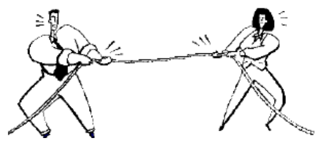
- Scenario…
  - You wish to compare two input devices for remote pointing (e.g., at a projection screen)
- External validity is improved if the test environment mimics expected usage
- Test environment should probably…
  - Use a projection screen (not a CRT)
  - Position participants at a significant distance from screen (rather than close up)
  - Have participants stand (rather than sit)
  - Include an audience!
- But… is internal validity compromised?

York University – Department of Computer Science and Engineering

Experimental Procedure Example

- Scenario…
  - You wish to compare two text entry techniques for mobile devices
- External validity is improved if the experimental procedure mimics expected usage
- Test procedure should probably require participants to…
  - Enter representative samples of text (e.g., phrases containing letters, numbers, punctuation, etc.)
  - Edit and correct mistakes as they would normally
- But… is internal validity compromised?

The Tradeoff



- There is tension between internal and external validity
- The more the test environment and experimental procedures are "relaxed" (to mimic real-world situations), the more the experiment is susceptible to uncontrolled sources of variation, such as pondering, distractions, or secondary tasks

Strive for the Best of Both Worlds

- Internal and external validity are increased by…
  - Posing multiple narrow (***testable***) questions that cover the range of outcomes influencing the broader (***untestable***) questions
    - E.g., a technique that is ***faster***, is ***more accurate***, takes ***fewer steps***, is ***easy to learn***, and is ***easy to remember***, is generally ***better***
- The good news
  - There is usually a positive correlation between the ***testable*** and ***untestable*** questions
    - I.e., participants generally find a UI ***better*** if it is ***faster***, ***more accurate***, ***takes fewer steps***, etc.

The Siren Call of the Skeptic

- There’s a *gotcha* in the previous slide
- The “good news” means we don’t need empirical research
- We just do a user study and ask participants which technique they preferred
- Because of the “positive correlation”, we needn’t waste our time on all this *gobbly-gook* data collection and analysis

Better[1] vs Better[2]

- A few points…
  - If participants are asked which technique they prefer, they’ll probably give an answer… even if they really have no particular preference! (There are many reasons, such as how recently they were tested on a technique, personal interaction with the investigator, etc.)
  - How much better? (A new technique might be deemed worthwhile only if the performance improvement is greater than, say, 20%.)
  - What are the strengths, weaknesses, limits, capabilities of the technique? (Are there opportunities to improve the technique?)
  - We need ***measurements*** to answer these questions!

[1] Aggregate outcome of answers to narrow (testable) empirical questions
[2] Answer to broad (untestable) questions

Answering Empirical Questions

- We want to know if the measured performance on a dependent variable (e.g., speed) is different between test conditions, so…
  - We conduct a user study and measure the performance on each test condition over a group of participants
  - For each test condition we compute the mean score over the group of participants
  - Then what?

Next slide

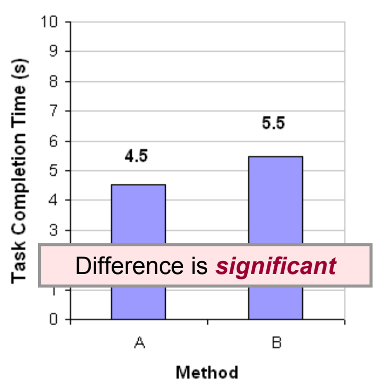
Answering Empirical Questions (2)

- Three questions:
  1. Is there a difference?
  2. Is the difference large or small?
  3. Is the difference significant or is it due to chance?
- Question #1 – obvious (some difference is likely)
- Question #2 – statistics can't help (Is a difference of 5% large or small?)
- Question #3 – statistics can help
- The basic statistical tool for Question #3 is the analysis of variance (anova)

Analysis of Variance

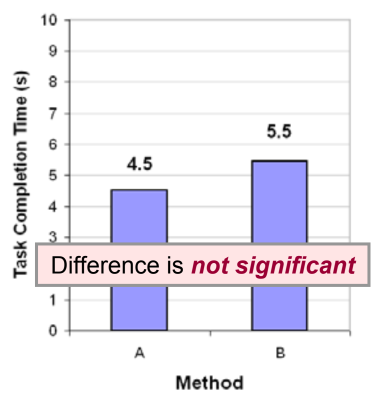
- It is interesting that the test is called an analysis of ***variance***, yet it is used to determine if there is a significant difference between the ***means***.
- How is this?

Example #1



"Significant" implies that in all likelihood the difference observed is due to the test conditions (Method A vs. Method B).

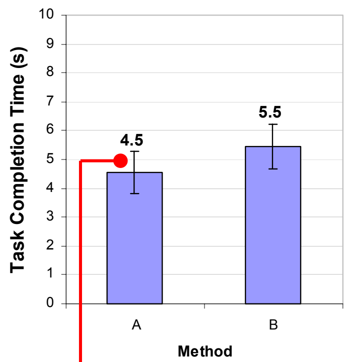
Example #2



"Not significant" implies that the difference observed is likely due to chance.

File: AnovaDemo.xls

Example #1 - Details



Example #1		
Participant	Method	
	A	B
1	5.3	5.7
2	3.6	4.6
3	5.2	5.1
4	3.3	4.5
5	4.6	6.0
6	4.1	7.0
7	4.0	6.0
8	5.0	4.6
9	5.2	5.5
10	5.1	5.6
Mean	4.5	5.5
SD	0.73	0.78

Error bars show
±1 standard deviation

Note: *SD* is the square root of the variance

Example #1 - Anova

StatView

ANOVA Table for Speed

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.839	.649				
Method	1	4.161	4.161	8.443	.0174	8.443	.741
Method * Subject	9	4.435	.493				

Probability that the difference in the means is due to chance

Reported as…

$F_{1,9} = 8.443$, $p < .05$

Thresholds for "p"
- .05
- .01
- .005
- .001
- .0005
- .0001

How to Report an F-statistic

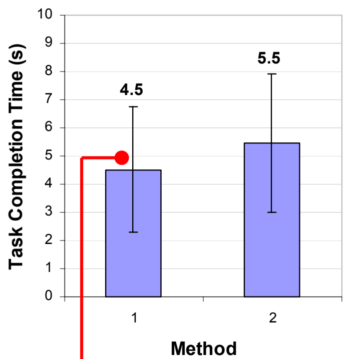
There was a significant main effect of input method on entry speed ($F_{1,9}$ = 8.44, p < .05).

- Notice in the parentheses
  - Uppercase for *F*
  - Lowercase for *p*
  - Italics for *F* and *p*
  - Space both sides of equal sign
  - Space after comma
  - Space both sides of less than sign
  - Degrees of freedom are subscript, plain, smaller font
  - Three significant figures for *F* statistic
  - No zero before the decimal point in the *p* statistic (except in Europe)

Or…

The mean task completion time for Method A was 4.5 seconds. This is 20.1% less than the mean of 5.5 seconds observed for Method B. The difference was statistically significant ($F_{1,9}$ = 8.44, p < .05).

Example #2 - Details



Example #2		
Participant	Method	
	A	B
1	2.4	6.9
2	2.7	7.2
3	3.4	2.6
4	6.1	1.8
5	6.4	7.8
6	5.4	9.2
7	7.9	4.4
8	1.2	6.6
9	3.0	4.8
10	6.6	3.1
Mean	4.5	5.5
SD	2.23	2.45

Error bars show
±1 standard deviation

Example #2 – Anova

StatView

ANOVA Table for Speed

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Pow er
Subject	9	37.017	4.113				
Method	1	4.376	4.376	.634	.4462	.634	.107
Method * Subject	9	62.079	6.898				

Probability that the difference in the means is due to chance

Reported as…

$F_{1,9}$ = 0.634, ns

Note: For non-significant effects, use "ns" if $F < 1.0$, or "$p > .05$" if $F > 1.0$.

And…

The mean task completion times were 4.5 seconds for Method A and 5.5 seconds for Method B. As there was substantial variation in the observations across participants, the difference proved not statistically significant as revealed in an analysis of variance ($F_{1,9}$ = 0.626, ns).

York University – Department of Computer Science and Engineering

StatView[a] Demo

Files:

`AnovaExample1.svd`
`AnovaExample2.svd`

[a] Now sold as JMP (see `http://www.statview.com`)

Anova Demo – *Anova2* †

Files:
`AnovaExample1.txt`
`AnovaExample2.txt`

```
DOS
CHI2007>java Anova2 AnovaExample1.txt 10 2 . . -a

==========================================================================
Effect                   df         SS             MS          F        p
--------------------------------------------------------------------------
Participant               9         5.940          0.660
F1                        1         4.232          4.232       8.449    0.01740
F1_x_Par                  9         4.508          0.501
==========================================================================
```

† This program and its API are available free to attendees of this course

Scientific Method (classical view)

Hypothesis = research question

Phenomenon = an interaction between a human and a computer (technology)

- Four steps…
  1. Observe and describe a ***phenomenon***
  2. Formulate an ***hypothesis*** to explain it
  3. Use the hypothesis to ***predict*** or ***describe*** ***other phenomena***
  4. Perform ***experiment*** to test hypothesis

Very nice, but what to researchers actually do?

next slide

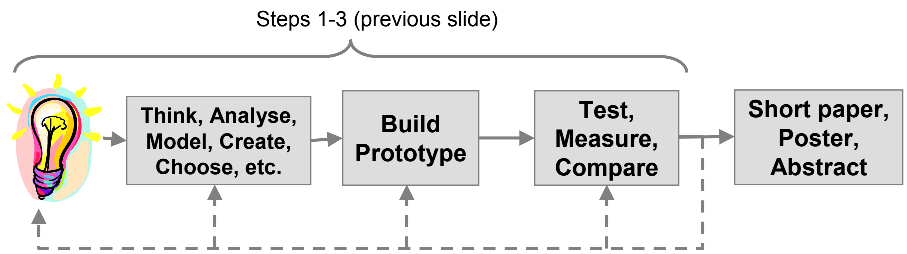
Other interactions

Experiment = user study

Predict = predictive model
Describe = descriptive model

Steps in Empirical Research (1)

Phase I – The Prototype



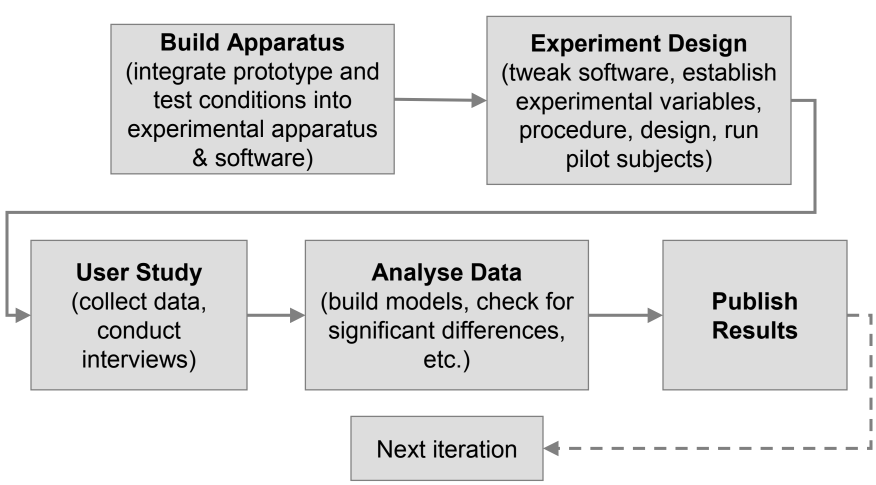
Iterations are frequent, unstructured, intuitive, informed, …

Research questions "take shape" (I.e., certain measurable aspects of the interaction suggest "test conditions", and "tasks" for empirical inquiry.

York University – Department of Computer Science and Engineering

Steps in Empirical Research (2)

Phase II – The User Study



Experiment Design

- Experiment design is a general term referring to the organization of variables, procedures, etc., in an experiment
- The process of designing an experiment is the process of deciding on which variables to use, what procedure to use, how many participant to use and how to solicit them, etc.
- Let's begin with some terminology…

Experiment Design - Terminology

- Terms to know
  - Participant
  - Independent variable (test conditions)
  - Dependent variable
  - Control variable
  - Random variable
  - Confounding variable
  - Within subjects vs. between subjects
  - Counterbalancing
  - Latin square

Participant

- The people participating in an experiment are referred to as participants
- Previously the term subjects was used, but it is no longer in vogue
- When referring specifically to the experiment, use the term participants (e.g., "all participants exhibited a high error rate…")
- General comments on the problem or conclusions drawn from the results may use other terms (e.g., "these results suggest that users are less likely to…"

York University – Department of Computer Science and Engineering

Independent Variable

- An independent variable is a variable that is selected or controlled through the design of the experiment
- Examples include device, feedback mode, button layout, visual layout, gender, age, expertise, etc.
- The terms independent variable and factor are synonymous

Test Conditions

- The levels, values, or settings for an independent variable are the test conditions
- Provide names for both a independent variable (factor) and the test conditions (levels) for the controlled variable
- Examples

Factor	*Levels ("test conditions")*
Device	mouse, trackball, joystick
Feedback mode	audio, tactile, none
Task	pointing, dragging
Visualization	2D, 3D, animated

Dependent Variable

- A dependent variable is a variable representing the measurements or observations on a independent variable
- Examples include task completion time, speed, accuracy, error rate, throughput, target re-entries, retries, key actions, etc.
- Provide a name for both the dependent variable and its units
- Examples:
  - Task completion time (ms), speed (word per minute, selections per minute, etc), error rate (%), throughput (bits/s), target re-entries (count, count per trial, etc.)

Control Variable

- Circumstances or factors that (a) might influence a dependent variable, but (b) are not under investigation need to be accommodated in some manner
- One way is to control them – to treat them as control variables
- E.g., room lighting, background noise, temperature
- The disadvantage to having too many control variables is that the experiment becomes less generalizable (i.e., applicable to other situations)

Random Variable

- Instead of controlling all circumstances or factors, some might be allowed to vary randomly
- Such circumstances are random variables
- More variability is introduced in the measures (that's bad!), but the results are more generalizable (that's good!)

Confounding Variable

- Any variable that varies systematically with an independent variable is a confounding variable
- E.g., if three devices are always administered in the same order, participant performance might improve doe to practice; I.e., from the 1^{st} to the 2^{nd} to the 3^{rd} condition; thus "practice" is a confounding variable (because it varies systematically with "device")

York University – Department of Computer Science and Engineering

Within Subjects, Between Subjects

- The administering of levels of a factor is either within subjects or between subjects
- If each participant is tested on each level, the factor is within subjects
- If each participant is tested on only one level, the factor is between subjects. In this case a separate group of participants is used for each condition.
- The terms repeated measures and within subjects are synonymous.

Within vs. Between Subjects

- Question: In designing an experiment, is it best to assign factors within subjects or between subject?
- Answer: It depends!
- Sometimes a factor must be between subjects (e.g., gender, age)
- Sometimes a factor must be within subjects (e.g., session, block)
- Sometimes there is a choice. In this case there is a tradeoff
- Within subjects advantage: (i) the variance due to participants' pre-dispositions should be the same across test conditions (cf. between subjects) (ii) simpler, because there are fewer participants
- Between subjects advantage: avoids interference effects (e.g., typing on two different layouts of keyboards)

Counterbalancing

- For repeated measures designs, participants' performance may tend to improve with practice as they progress from one test condition to the next. Thus, participants may perform better on the second condition simply because they benefited from practice on the first. This is bad news.
- To compensate, the order of presenting conditions is counterbalanced
- Participants are divided into groups, and a different order of administration is used for each group
- The order is best governed by a Latin Square (next slide)

Latin Square

- The defining characteristic of a Latin Square is that each condition occurs only once in each row and column
- Examples:

3 X 3 Latin Square

A	B	C
B	C	A
C	A	B

4 x 4 Latin Square

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

4 x 4 Balanced Latin Square

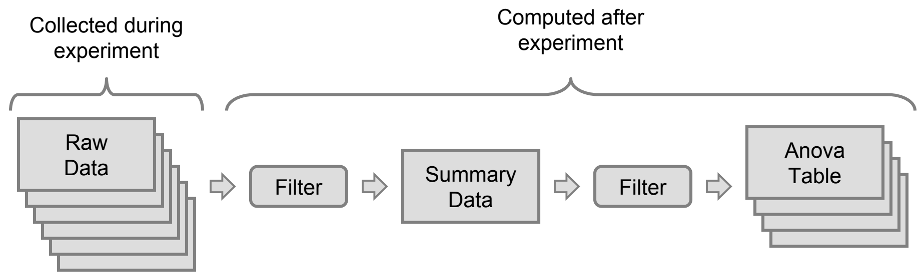
A	B	C	D
B	D	A	C
D	C	B	A
C	A	D	B

Note: In a balanced Latin Square each condition both precedes and follows each other condition an equal number of times

Succinct Statement of Design

- "3 x 2 repeated-measures design" refers to an experiment with two factors, having three levels on the first, and two levels on the second. There are six test conditions in total. Both factors are repeated measures, meaning all participants were tested on all test conditions
- Note: A mixed design is also possible
  - In this case, the levels for one factor are administered to all participants (within subjects) while the levels for another factor are administered to separate groups of participants (between subjects).

Data Collection and Progression



Raw Data

- Vertical format, unstructured, ugly!
- Primarily contains timestamps and events
- Also identifies test conditions, participant, session, etc., either in filename or within file (needed later)
- Advice: plan ahead!

Summary Data

- Rectangular
- One row per "unit of analysis"
- Formatted for importing into spreadsheet
- Columns for test conditions (largely redundant, but useful), and dependent measures (aggregated as per the "unit of analysis")

Anova Table

- Rectangular
- One row per participant
- Formatted for importing into stats package
- Cells contain dependent measures
- One file per dependent measure

York University – Department of Computer Science and Engineering

Case Study

- Scenario…

Researcher R has an interest in the application of eye tracking technology to the problem of text entry. After studying the existing body of research and commercial implementations, R develops some ideas on how to improve the interaction. R initiates a program of empirical inquiry to explore the performance limits and capabilities of various feedback modalities for keys in on-screen keyboards used with eye typing.

Reality check

FYI, see

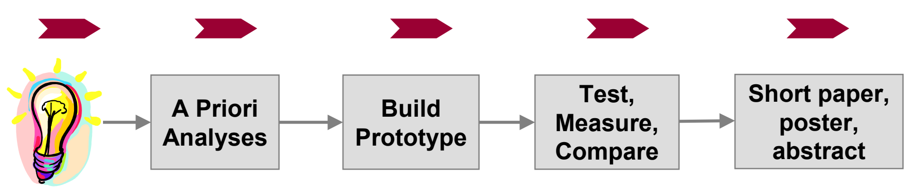
http://www.yorku.ca/mack/chi03d.html

57

York University – Department of Computer Science and Engineering

Case Study (reality check)

Phase I – The Prototype



58

The User Study (1)

- Participants
  - 13, volunteer, recruited from university campus, age, gender, computer experience, eye tracking/typing experience
- Apparatus
  - Describe hardware and software, etc.

The User Study (2)

- Experiment design
  - 4 x 4 repeated measures design
  - Independent variables (viz. factors)…
    - Feedback modality (A0, CV, SV, VO)
    - Block (1, 2, 3, 4)
  - Dependent variables (viz. measures)
    - Speed (in "words per minutes")
    - Accuracy (in "percentage of characters in error")
    - Key activity (in "keystrokes per character")
    - Eye activity (in "read presented text events per phrase")
    - Etc. (other "events" of interest)
    - Also… responses to "broad" questions
  - Order of conditions
    - Feedback modality order differed for each participant

The User Study (3)

- Procedure
  - General objectives of experiment explained
  - Eye tracking apparatus calibrated
  - Practice trials, then
  - Data collection begins
  - Phrases of text presented by experimental software
  - Participants instructed to enter phrases "as quickly and accurately as possible"
  - Five phrases entered per block
  - Total number of phrases entered in experiment…
    - 13 x 4 x 4 x 5 = 1040

Experiment Replication

- The description of the experimental methodology (i.e., participants, participant selection, apparatus, design, procedure) must be sufficient to allow the experiment to be replicated by other researchers
- This is necessary to allow the possibility for the results to be verified or refuted
- An experiment that cannot be replicated is useless and does not merit publication

User Study (4)

- Raw data (208 files)
  - `CaseStudy-RawData.txt`
- Summary data (1 file)
  - `CaseStudy-SummaryData.txt`
- Anova Tables (~5 files)
  - `CaseStudy-AnovaTable.txt`

View from editor

York University – Department of Computer Science and Engineering

The User Study (5)

- Results for speed (only example given here)
  - Grand mean = 6.96 wpm
  - By feedback modality…
  - By block…
  - Salient observations
    - 4th block speed for best condition was…

Anova Data Table (not in paper)

Factors and levels

Speed																	
	A	A	A	A	C	C	C	C	S	S	S	S	V	V	V	V	
Participant	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	Mean
1	6.17	7.19	7.04	7.09	6.76	7.40	7.54	7.94	6.44	6.17	7.84	6.81	5.20	6.29	7.39	7.63	6.93
2	6.71	7.25	7.05	7.15	7.73	7.57	8.04	7.26	7.00	6.75	7.68	7.46	7.50	7.07	7.32	7.06	7.29
3	6.80	6.65	7.62	7.98	6.61	7.18	7.34	8.19	6.65	7.53	7.09	7.90	5.73	7.24	6.94	7.13	7.16
5	6.30	6.31	7.59	7.38	6.85	7.64	7.58	7.88	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6.59	6.97	7.72	7.12
7	6.68	6.89	7.32	7.51	7.00	7.81	7.64	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	7.57	7.20	7.11
8	6.08	6.55	6.83	5.92	7.44	6.93	7.56	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	7.45	7.16	6.98
9	7.62	7.01	6.60	7.07	6.91	6.81	6.91	7.73	6.50	7.57	7.59	7.80	6.62	7.06	7.16	7.41	7.15
10	5.88	5.71	7.33	7.11	6.66	7.97	7.64	8.15	6.35	7.21	6.56	7.33	5.00	6.97	6.54	6.36	6.80
12	6.89	7.61	7.42	7.88	7.79	8.28	8.20	8.39	6.62	6.87	7.99	8.23	9.57	8.17	7.91	7.09	7.81
13	6.85	6.57	8.14	6.00	5.92	7.89	7.49	6.98	6.05	7.45	5.34	7.46	7.21	6.81	6.80	8.24	6.95
14	5.37	5.56	6.04	6.86	6.20	6.82	7.71	7.76	5.85	6.37	6.74	6.69	5.98	6.43	6.38	5.87	6.41
15	5.51	6.12	6.32	7.00	6.16	6.49	7.21	7.19	5.65	6.52	6.49	7.10	5.31	6.88	6.36	6.93	6.45
16	5.88	7.18	5.95	6.00	4.85	6.98	7.37	6.98	6.88	6.21	4.96	5.34	6.72	7.14	4.96	6.80	6.26
																	6.96

Outlier (explain in paper)

Each cell is the mean for five phrases of input

Anova Table (not in paper)

ANOVA Table for Entry Speed (wpm)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	12	32.319	2.693				
Feedback Mode	3	8.210	2.737	8.772	.0002	26.317	.994
Feedback Mode * Subject	36	11.231	.312				
Block	3	13.310	4.437	10.923	<.0001	32.768	.999
Block * Subject	36	14.623	.406				
Feedback Mode * Block	9	1.772	.197	.633	.7669	5.694	.294
Feedback Mode * Block * Subject	108	33.606	.311				

Verbal statement and discussion of findings will include…

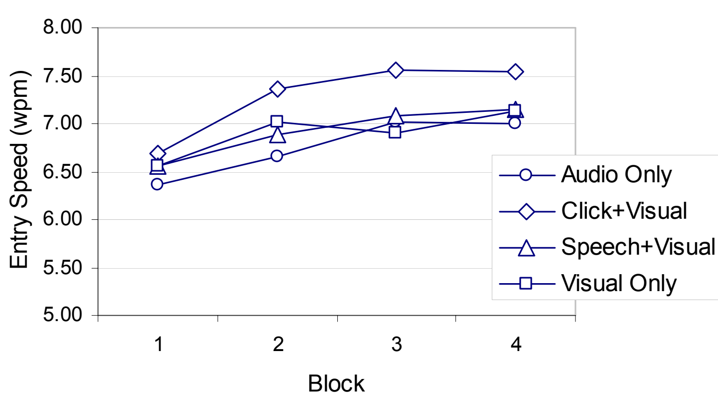
- Main effect for Feedback mode ***significant***: $F_{3,36} = 8.77$, $p < .0005$
- Main effect for Block ***significant***: $F_{3,36} = 10.92$, $p < .0001$
- Feedback mode by block interaction ***not significant***: $F_{9,108} = 0.767$, ns

Summary Table for Speed (not in paper)

Speed (wpm)					
	Feedback Mode				
Block	Audio Only	Click+Visual	Speech+Visual	Visual Only	mean
1	6.36	6.68	6.56	6.55	6.54
2	6.66	7.37	6.88	7.02	6.98
3	7.02	7.56	7.09	6.90	7.14
4	7.00	7.55	7.14	7.12	7.20
mean	6.76	7.29	6.92	6.90	6.97

5.7% faster on 4th block

Summary Chart (in paper!)

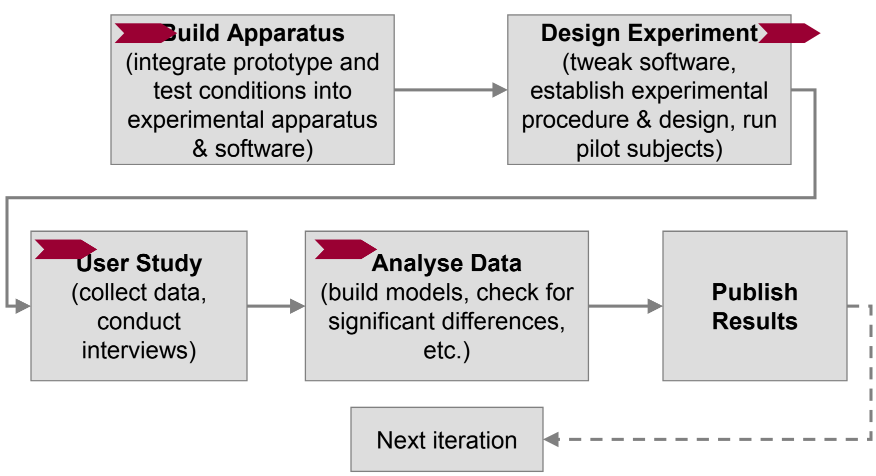


The Broad Questions

- Participants were asked to rank the feedback mode based on personal preference
- Results
  - Six of 13 participants gave a 1st place ranking to the fastest feedback modality
    - Not a strong result
    - Probably the differences just weren't large enough for participants to really tell the difference in overall performance.
  - Notably, ten of 13 participants gave a 1st or 2nd place ranking to the fastest feedback modality
    - Thus, there is a modest trend that better performance yields a better preference rating (but empirical research is the key!)

Case Study (reality check)

Phase II – The User Study



What's Missing?

- The case study just described is interesting, but something is missing
- There is no…
  - theoretical account of the phenomena
- There is no…
  - delineation, description, categorization of the known and observed behaviors (…that can form such a theoretical account)

Empirical Research in HCI
Why? (The Second Answer)

While these empirical results are of direct use in selecting an ***Interaction technique****,*[1] *it would obviously be of greater benefit if a theoretical account of the results could be made. For one thing, the need for some experiments might be obviated; for another, ways of improving* ***interaction***[1] *might be suggested.*

Card, English, and Burr (1978, p. 608)

- Why… to build and test ***models of interaction***

[1] Edited to recast in general terms

Case Study: The Case for a Model

- Is there a "model of interaction" suggested by the observations in the case study?
- Perhaps. Here's one possibility
- All gaze point changes were logged as "events"
- What was the total number of such events?
- Are there categories of such events?
- The identification, labeling, and tabulation of such could form the basis of a model of interaction for eye typing

York University – Department of Computer Science and Engineering

Thank you

Questions?

Suggested Readings

1. Card, S. K., English, W. K., and Burr, B. J. Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT, *Ergonomics 21* (1978), 601-613.
2. Carroll, J. M. (ed.), *Toward a multidisciplinary science of human-computer interaction*, (San Francisco: Morgan Kaufmann, 2003).
3. Kaindl, H. Methods and modeling: Fiction or useful reality?, *Extended Abstracts of CHI 2001*. (2001), 213-214.
4. Newell, A., and Card, S. K. The prospects for psychological science in human-computer interaction, *Human-Computer Interaction 1* (1985), 209-242.