Empirical Research Methods for Human-Computer Interaction

I. Scott MacKenzie

York University
Toronto, Canada

http://www.yorku.ca/mack

Sponsors

Presenter

Scott MacKenzie's research is in human-computer interaction with an emphasis on human performance measurement and modeling, experimental methods and evaluation, interaction devices and techniques, text entry, touch-based input, language modeling, accessible computing, gaming, and mobile computing. He has more than 160 peer-reviewed publications in the field of Human-Computer Interaction (including more than 30 from the ACM's annual SIGCHI conference) and has given numerous invited talks over the past 25 years. In 2015, he was elected into the ACM SIGCHI Academy. That same year he was the recipient of the Canadian Human-Computer Communication Society's (CHCCS) Achievement Award. Since 1999, he has been Associate Professor of Computer Science and Engineering at York University, Canada.

Home page: http://www.yorku.ca/mack/

Recent book: http://www.yorku.ca/mack/HCIbook/

2

Topics

- The what, why, and how
- Group participation in a real experiment
- Observations and measurements
- Research methods (and their properties)
- Experiment terminology
- Experiment design
- ANOVA statistics and experiment results
- Parts of a research paper

3

What is…

Research

(three dictionary definitions)

1. Careful or diligent search

2. Collecting information about a subject

3. Investigation or experimentation aimed at the discovery and interpretation of facts

4

Definition #4
(not in dictionary)

5

Definition #4
(not in dictionary)

- Research → *a word added to give weight to baseless assertions intended to deceive the public*

6

Example (Definition #4)

- *"Independent research proves our Internet service is the fastest and most reliable – period."*

Rogers Communications, Inc.

7

Example (Definition #4)

- *"Independent research proves our Internet service is the fastest and most reliable – period."*

Rogers Communications, Inc.

8

Example (Definition #4)

- *"Independent research proves our Internet service is the fastest and most reliable – period."*

Rogers Communications, Inc.

9

What is…

Empirical research

- cf. theoretical research
- Properties of empirical research:
  - Based on observation or experience
  - Relying on observation or experience alone without due regard for system or theory (i.e., not blinded by pre-conceptions)
  - Capable of verification or disproval by observation or experiment
- In HCI…
  - "observation or experience" is of humans interacting with computers (or technology of some sort)

10

What is…

Empirical Research

- cf. theoretical research
- Properties of empirical research:
  - Based on observation or experience
  - Relying on observation or experience alone without due regard for system or theory (i.e., not blinded by pre-conceptions)
  - Capable of verification or disproval by observation or experiment
- In HCI…
  - "observation or experience" is of humans interacting with computers (or technology of some sort)

11

Why do…

Empirical Research

- We conduct empirical research to…
  - Answer (and raise!) questions about new or existing user interface designs or interaction techniques
  - Find *cause-and-effect* relationships
  - Transform baseless opinions into informed opinions supported by evidence
  - Develop or test models that *describe* or *predict* behavior (of humans interacting with computers)

12

How do we do…

Empirical Research

- Through a program of inquiry conforming to the ***scientific method***
- The scientific method involves…
  - The recognition and formulation of a problem
  - The formulation and testing of hypotheses
  - The collection of data through observation and experiment
- In HCI…
  - The methodology is often a ***user study*** (an experiment with human participants)

13

Topics

- The what, why, and how
- Group participation in a real experiment
- Observations and measurements
- Research methods (and their properties)
- Experiment terminology
- Experiment design
- ANOVA statistics and experiment results
- Parts of a research paper

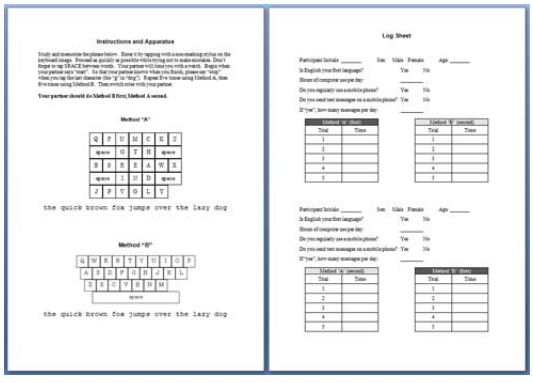
14

Group Participation

- At this point in the course, attendees are divided into groups of two to participate in a real user study
- A two-page handout is distributed to each group (see next slide)
- Read the instructions on the first page and discuss the procedure with your partner
- The instructors will provide additional information

15

Handout (2 pages)



Full-size copies of the handout pages will be distributed during the course. The pages are also available on the course web site.

16

Do the Experiment

- The experiment is performed
- This takes about 30 minutes
- Assistants transcribe the tabulated data into a ready-made spreadsheet
- Results are presented after the break, in Session Two

17

Topics

- The what, why, and how
- Group participation in a real experiment
- Observations and measurements
- Research methods (and their properties)
- Experiment terminology
- Experiment design
- ANOVA statistics and experiment results
- Parts of a research paper

18

Observations and Measurements

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

19

Scales of Measurement[1]

- Nominal
- Ordinal
- Interval
- Ratio

crude

sophisticated

[1] Stevens S.S. (1946, June 7). On the theory of scales of measurement. *Science*, pp. 677-680.

20

Scales of Measurement[1]

- Nominal
- Ordinal
- Interval
- Ratio

[1] Stevens S.S. (1946, June 7). On the theory of scales of measurement. *Science*, pp. 677-680.

21

Scales of Measurement

- ***Nominal***
- Ordinal
- Interval
- Ratio

- (aka ***categorical data***) – arbitrary codes assigned to attributes; e.g.,
  - M = male, F = female
  - 1 = audio feedback, 2 = vibrotactile feedback
- Stats:
  - equivalence, ~~greater/less than~~, ~~mean~~, ~~ratio~~
- Usually, it is the count that is important
  - "Are females or males more likely to…"
- Example:

Gender	Mobile Phone Usage		Total	%
	Not Using	Using		
Male	683	98	781	51.1%
Female	644	102	746	48.9%
Total	1327	200	1527	
%	86.9%	13.1%		

Note: The counts (grey) are ratio scale measurements

Real Data!

22

Scales of Measurement

- Nominal
- ***Ordinal***
- Interval
- Ratio

- Associates a rank to an attribute
- The attribute is any characteristic of interest, for example
  - Users try three different GPS systems, then rank them: 1st, 2nd, 3rd choice
- Stats:
  - equivalence, greater/less than, ~~mean~~, ~~ratio~~
- Example:

What is your weekly time playing computer games?

1. 0 hr
2. 1 - 5 hr
3. 5 - 20 hr
4. 20 - 40 hr
5. More than 40 hr

23

Scales of Measurement

- Nominal
- Ordinal
- ***Interval***
- Ratio

- Equal distances between adjacent values
- No absolute zero (ratios not possible)
- Classic example: temperature (°F, °C)
- Stats:
  - equivalence, greater/less than, mean, ~~ratio~~
- Example: Likert scale questionnaire responses

Indicate your level of agreement with the following statement:

	Strongly disagree				Strongly agree
It is safe to talk on a mobile phone while driving.	1	2	3	4	5

24

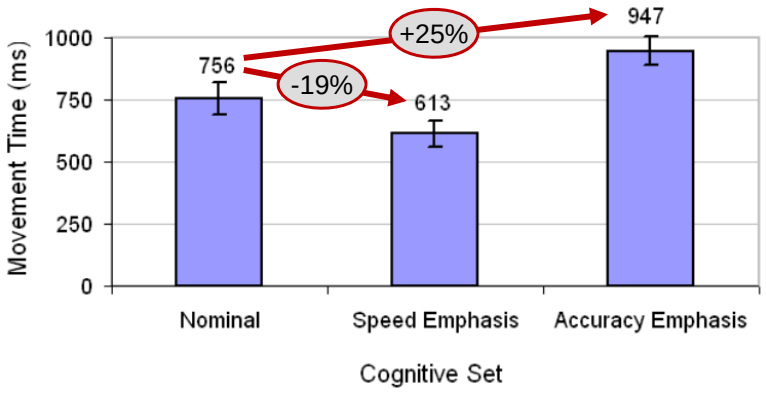
Scales of Measurement

- Nominal
- Ordinal
- Interval
- ***Ratio***

- (aka ***continuous data***) most sophisticated of the four scales of measurement
- Preferred scale of measurement
- Stats:
  - equivalence, greater/less than, mean, ratio
- Absolute zero, therefore many calculations possible
- Often, ratio data are counts; e.g.,
  - "time" – the number of seconds to complete a task
  - "DEL presses" – the number of times the delete key was pressed
- Example: (next slide)

25

Ratio Data Example in HCI[1]



$F_{2,34} = 372.7,\ p < .0001$

[1] MacKenzie, I. S., & Isokoski, P. (2008). Fitts' throughput and the speed-accuracy tradeoff. *Proc CHI 2008*, pp. 1633-1636.

26

Topics

- The what, why, and how
- Group participation in a real experiment
- Observations and measurements
- Research methods (and their properties)
- Experiment terminology
- Experiment design
- ANOVA statistics and experiment results
- Parts of a research paper

27

Allen Newell (1927-1992)

- ACM Turing Award (1975)
- With Stuart Card and Tom Moran, author of *The Psychology of Human-Computer Interaction* (1983)

"Science is method. Everything else is commentary."

28

Research Methods

Theoretical Research

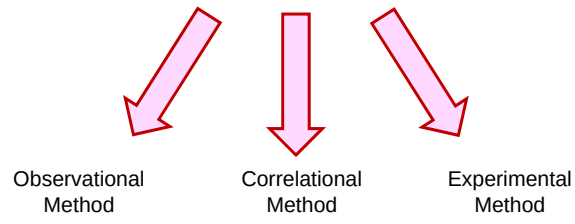
Empirical Research

29

Research Methods

Theoretical Research

Empirical Research

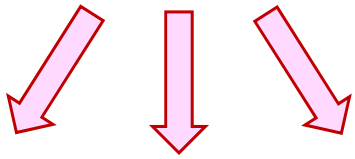


30

Research Methods

Theoretical Research

Empirical Research



Observational Method

Correlational Method

Experimental Method

31

Observational Method

- Example techniques:
  - Interviews, field investigations, contextual inquiries, case studies, field studies, focus groups, think aloud protocols, story telling, walkthroughs, cultural probes, etc.
- Focus on ***qualitative*** assessments (cf. quantitative)
- Relevance vs. precision
  - Higher in relevance (behaviours studied in a natural setting)
  - Lower in precision (lacks control available in a laboratory)
- Goal: discover and explain reasons underlying human behaviour (*why* or *how*, as opposed to *what*, *where*, or *when*)

32

Experimental Method

- Controlled experiment conducted in lab setting
- In HCI, this is typically called a ***user study***
- Focus on ***quantitative*** assessments (cf. qualitative)
- Relevance vs. precision
  - Lower in relevance (artificial environment)
  - Higher in precision (extraneous behaviours easy to control)
- At least two variables:
  - ***Manipulated variable*** *(*aka ***independent variable****)*
  - ***Response variable*** (aka ***dependent variable***)
- Cause-and-effect conclusions possible

33

Correlational Method

- Look for relationships between variables
- Observations made, data collected
  - Example: *Are user's privacy settings while social networking related to their age, gender, level of education, employment status, income, number of tattoos, etc.*
- Non-experimental
  - Interviews, on-line surveys, questionnaires, etc.
- Balance between relevance and precision
- Predictions possible
- Cause-and-effect conclusions not possible

34

Research Methods

Observational

- Real-world setting
- No variables per se
- Broad, qualitative questions:
  - What's going on?
- High-level inquiry

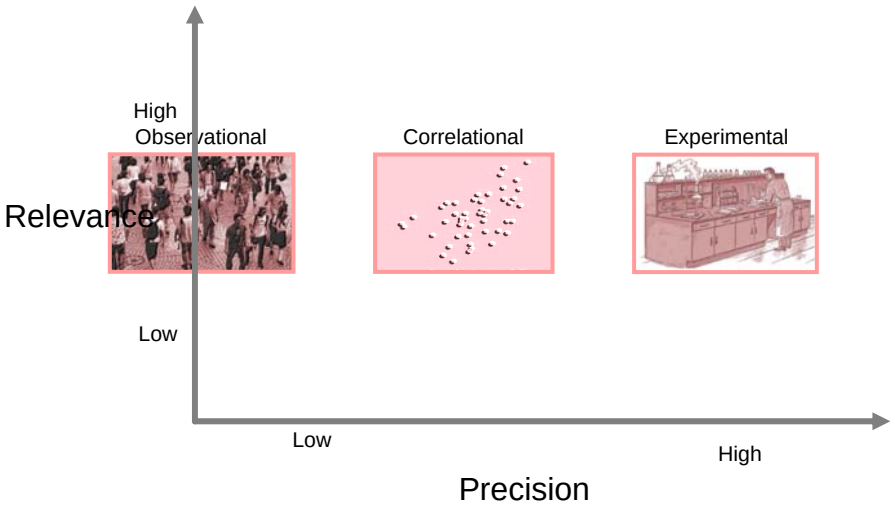
Correlational

Experimental

- Controlled setting (lab)
- IVs, DVs, etc.
- Narrow, quantitative questions:
  - How fast? How accurate?
- Low-level inquiry

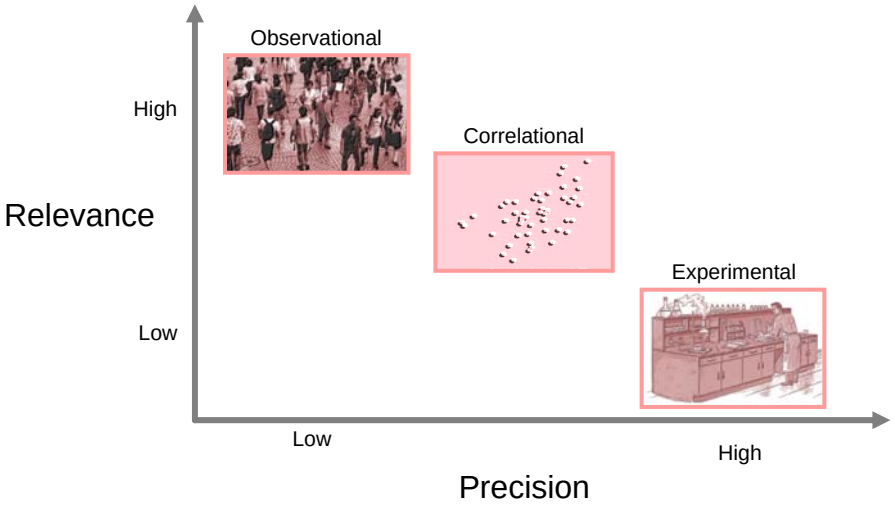
35

Relevance vs. Precision



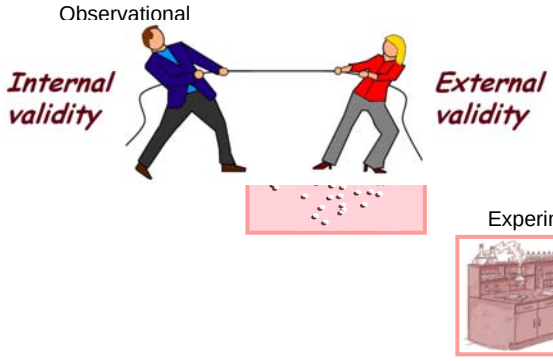
36

Relevance vs. Precision



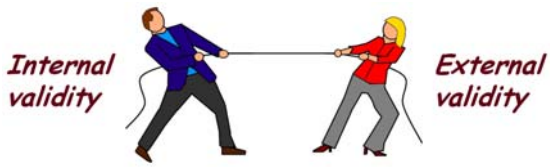
37

Internal Validity ↔ External Validity



38

Internal Validity ↔ External Validity



Experimental

Correlational

Observational

39

Internal Validity

- Definition:
  - The extent to which the effects observed are due to the ***test conditions*** (e.g., Method A vs. Method B)
- High internal validity means…
  - Differences (in the means) are due to ***inherent properties*** that distinguish the test conditions
  - Variances are due to ***participant pre-dispositions***, or are controlled or exist equally across the test conditions

40

External Validity

- Definition:
  - The extent to which results are generalizable to other ***people*** and other ***situations***
- High external validity means…
  - Participants are ***representative*** of the broader intended population of users
  - The ***test environment*** and ***experimental procedures*** are representative of real world situations where the interface or technique will be used

41

Research Questions

- Consider the following questions about a novel idea, perhaps a new interaction technique
  - Is it any good?
  - Is it better than current practice?
  - Which design or engineering alternative is better?
  - What are the performance limits?
  - What are the weaknesses?
  - Does it work well for novices?
  - How much practice is required?
- These questions, while unquestionably relevant, are not testable

42

Testable Research Questions

- Try to re-think your questions as testable questions (even though the new question may appear less important)
- Scenario…
  - You have an idea for a *new* text entry technique for mobile phones, and you think it's pretty neat, perhaps better than the existing Qwerty soft keyboard (QSK)
- Testable research question:
  - *Is the measured entry speed (in words per minute) higher for the new technique than for QSK after one hour of use?*

43

Topics

- The what, why, and how
- Group participation in a real experiment
- Observations and measurements
- Research methods (and their properties)
- Experiment terminology
- Experiment design
- ANOVA statistics and experiment results
- Parts of a research paper

44

Experiment Terminology (Part 1)

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

45

Participant

- The people participating in an experiment are referred to as ***participants*** (the term ***subjects*** is also acceptable[1])
- When referring specifically to the experiment, use ***participants***
  - *"all **participants** exhibited a high error rate…"*
- When discussing the problem generally or drawing conclusions, use other terms
  - *"these results suggest that **users** are less likely to…"*
- Report the selection criteria and give relevant demographic information or prior experience

[1] APA. (2010). *Publication Manual of the American Psychological Association* (6th ed.) Washington, DC: APA, p. 73.

46

How Many Participants

- Use the same number of participants as used in similar research[1]
- Too many participants…
  - and you get statistically significant results for differences of no *practical* significance
- Too few participants…
  - and you fail to get statistically significant results when there really is an inherent difference between the test conditions

[1] Martin D.W. (2004). *Doing psychology experiments* (6th ed.). Belmon, CA: Wadsworth, p. 234.

47

Independent Variable

- ***Independent variable*** – a circumstance that is manipulated through the design of the experiment
- It is "independent" because it is independent of participant behavior (i.e., there is nothing a participant can do to influence an independent variable)
- Examples
  - Interface, device, feedback mode, button layout, visual layout, gender, age, expertise, etc.
- The terms ***independent variable*** and ***factor*** are synonymous

48

Test Conditions

- The levels, values, or settings for an independent variable are the ***test conditions***
- Provide a names for both the ***independent variable*** and its ***levels*** (***test conditions***)
- Use these names consistently throughout a research paper
- Examples

Independent Variable	*Test Conditions (Levels)*
Device	mouse, touchpad, pointing stick
Feedback mode	audio, tactile, none
Task	pointing, dragging
Visualization	2D, 3D, animated
Search interface	Google, Bing

49

Dependent Variable

- ***Dependent variable*** – a measurable aspect of the interaction involving an independent variable
- Examples
  - Task completion time, speed, accuracy, error rate, throughput, target re-entries, task retries, presses of backspace, etc.
- Give a name to the dependent variable, separate from its units, for example…
  - "entry speed" in "words per minute"
  - "task completion time" in "seconds"
- Clearly define all dependent variables (research must be reproducible!)

50

Control Variable

- ***Control variable*** – a circumstance (not under investigation) that is held constant
- Upside: aides internal validity (better chance of obtaining statistical significance)
- Downside: hinders external validity (results are less generalizable to other people and other situations)
- Example: Consider an experiment on the effect of font color and background color on reader comprehension
  - Independent variables: font color, background color
  - Dependent variables: comprehension test scores
  - Control variables:
    - Font size (e.g., 12 point)
    - Font family (e.g., Times)
    - Ambient lighting (e.g., fluorescent, fixed intensity)

51

Random Variable

- ***Random variable*** – a circumstance that is allowed to vary randomly
- Upside: aides external validity (results are more generalizable)
- Downside: hinders internal validity (more variability is introduced in the measures)
- Example research question: *Does user stance affect performance while playing Guitar Hero?*
  - Independent variable: stance (standing, sitting)
  - Dependent variable: score on songs
  - Random variables
    - Prior experience playing a real musical instrument
    - Prior experience playing *Guitar Hero*
    - Amount of coffee consumed prior to testing

52

Tradeoff

(control variable vs. random variable)

- There is a trade-off which can be examined in terms of internal validity and external validity (see below)

Variable	Advantage	Disadvantage
Random	Improves external validity by using a variety of situations and people.	Compromises internal validity by introducing additional variability in the measured behaviours.
Control	Improves internal validity since variability due to a controlled circumstance is eliminated	Compromises external validity by limiting responses to specific situations and people.

53

Confounding Variable

- ***Confounding variable*** – a circumstance that varies systematically with an independent variable
- Upside: none!
- Downside: results misleading, even wrong
- Example: a study investigates "camera distance" in an eye tracking task
  - Independent variable: *camera distance* with levels *near* and *far*
    - Near setup: small camera mounted on eye glasses
    - Far setup: commercial eye tracker mounted below display
  - *Hardware* is a confounding variable
  - Are the differences observed due to camera distance or to the different hardware or software drivers?
  - No reliable conclusions are possible

54

Topics

- The what, why, and how
- Group participation in a real experiment
- Observations and measurements
- Research methods (and their properties)
- Experiment terminology
- Experiment design
- ANOVA statistics and experiment results
- Parts of a research paper

55

Experiment Design

- ***Experiment design*** – the process of deciding
  - What variables to use
  - What tasks and procedures to use
  - How many participants to use and how to solicit them
  - Etc.
- Let's continue with some terminology…

56

Experiment Terminology (Part 2)

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

57

Within-subjects, Between-subjects

- Two ways to assign conditions to participants:
  - ***Within-subjects*** → each participant is tested on each condition (aka *repeated measures*)
  - ***Between-subjects*** → each participant is tested on one condition only
  - Examples:

Within-subjects

Participant	Test Condition		
1	A	B	C
2	A	B	C

Between-subjects

Participant	Test Condition
1	A
2	A
3	B
4	B
5	C
6	C

58

Within-subjects

- Advantages
  - Fewer participants (easier to recruit, schedule, etc.)
  - Less "variation due to participants"
  - No need to balance groups (because there is only one group!)
- Disadvantage
  - Order effects (i.e., interference between conditions)

Between-subjects

- Disadvantages
  - More participants (harder to recruit, schedule, etc.)
  - More "variation due to participants"
  - Need to balance groups (to ensure they are more or less the same)
- Advantage
  - No order effects (i.e., no interference between conditions)

59

Within-subjects, Between-subjects (2)

- Sometimes…
  - A factor must be assigned within-subjects
    - Examples: block, session (if learning is the IV)
  - A factor must be assigned between-subjects
    - Examples: gender, handedness
  - There is a choice
    - In this case, the balance tips to within-subjects (see previous slide)
- With two factors, there are three possibilities:
  - both factors within-subjects
  - both factors between-subjects
  - one factor within-subjects + one factor between-subjects (this is a *mixed design*)

60

Counterbalancing

- Only applies to within-subjects designs:
  - Participants may benefit from the 1st condition and thereby perform better on the 2nd condition
  - This is a problem (results are misleading)
- To compensate, ***counterbalancing*** is used:
  - Participants are divided into ***groups***, and a different testing order is used for each group
- The testing order is best governed by a ***Latin Square*** (next slide)
- ***Group***, then, is a between-subjects factor
  - Was there an effect for group? Hopefully not!

61

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

62

Succinct Statement of Design

- "***3 x 2 within-subjects design***"
  - 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 conditions
- A mixed design is also possible
  - 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)

63

Topics

- The what, why, and how
- Group participation in a real experiment
- Observations and measurements
- Research methods (and their properties)
- Experiment terminology
- Experiment design
- ANOVA statistics and experiment results
- Parts of a research paper

64

Answering Research Questions

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

65

Answering Research Questions (2)

1. Is there a difference?
   - Some difference is likely
2. Is the difference large or small?
   - Statistics can't help (Is a 5% difference large or small?)
3. Is the difference of practical significance?
   - Statistics can't help (Is a 5% difference useful? People resist change!)
4. Is the difference real? (Is it statistically significant or is it due to chance?)
   - Statistics can help!
   - The statistical tool is the analysis of variance (ANOVA)

66

Null Hypothesis

- Formally speaking, a research question is not a question. It is a statement called the ***null hypothesis***.
- Example:

> There is no difference in entry speed between Method A and Method B.

- Assumption of "no difference"
- Research usually seeks to reject the null hypothesis
- Please bear in mind, with experimental research…
  - We gather and test evidence
  - We do not prove things

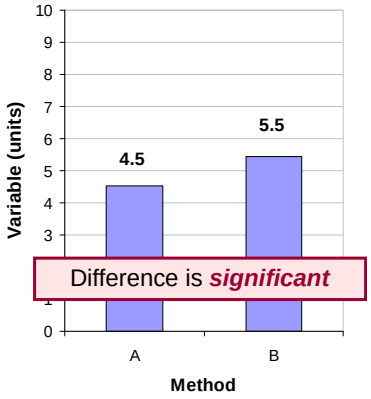
67

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?

68

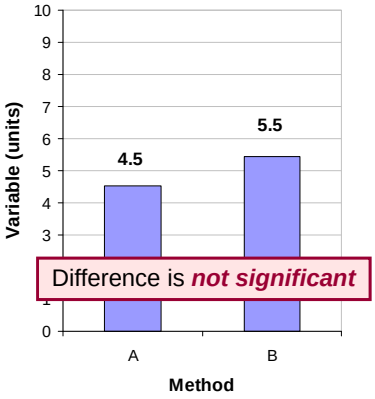
Example #1



Difference is ***significant***

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

Example #2

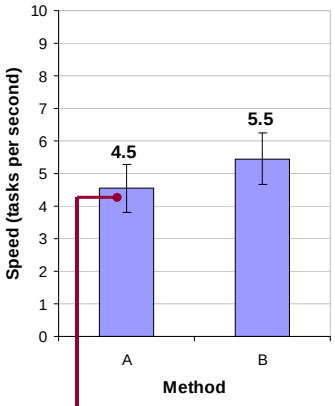


Difference is ***not significant***

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

69

Example #1 - Details



Example #1		
Participant	Method	
	A	B
1	5.3	5.7
2	3.6	4.8
3	5.2	5.1
4	3.5	4.5
5	4.6	6.0
6	4.1	6.8
7	4.0	6.0
8	4.8	4.6
9	5.2	5.5
10	5.1	5.6
Mean	4.5	5.5
SD	0.68	0.72

Error bars show ±1 standard deviation

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

70

Example #1 - ANOVA

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.080	.564				
Method	1	4.232	4.232	9.796	.0121	9.796	.804
Method * Subject	9	3.888	.432				

Probability of obtaining the observed data if the null hypothesis is true

Reported as…

$F_{1,9} = 9.796,\ p < .05$

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

71

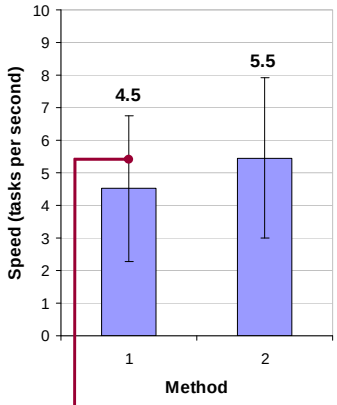
How to Report an *F*-statistic

There was a significant effect of input method on entry speed ($F_{1,9} = 9.796$, $p < .05$).

- 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 on both sides of less-than sign
  - Degrees of freedom are subscript, plain, smaller font
  - Three (maybe four) significant figures for *F* statistic
  - No zero before the decimal point in the *p* statistic

72

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

73

Example #2 – ANOVA

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	37.372	4.152				
Method	1	4.324	4.324	.626	.4491	.626	.107
Method * Subject	9	62.140	6.904				

Probability of obtaining the observed data if the null hypothesis is true

Reported as…

$F_{1,9} = 0.626$, ns

Two ways of reporting non-significant effects:

- If $F < 1.0$, use "ns"
- If $F > 1.0$, use "$p > .05$"

74

Reporting an *F*-statistic – Revisited

- Helpful to mention both the independent variable and the dependent variable:

> "The effect of ***independent_variable*** on ***dependent_variable*** was statistically significant (F-statistic)."

- Example on next slide

75

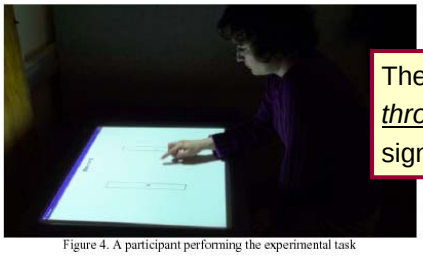


Figure 4. A participant performing the experimental task

RESULTS AND DISCUSSION

Throughput

Touch interaction yielded a higher throughput compared to the mouse. The overall mean throughput for touch interaction was 5.52 bps, which was 44.1% higher than the 3.83 bps observed for the mouse. The effect of input technique on throughput was statistically significant ($F_{1,11}$ = 35.51, p < .0001). Although not as high as the throughput reported by Forlines et al. (2007) for touch input (discussed earlier), our throughput values were computed using a direct

The effect of *input technique* on *throughput* was statistically significant ($F_{1,11}$ = 35.51, p < .0001).

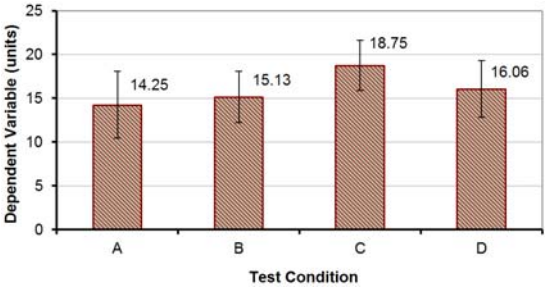
Independent variable:
Input technique

Dependent variable:
Throughput

Sasangohar, F., MacKenzie, I. S., & Scott, S. D. (2009). Evaluation of mouse and touch input for a tabletop display using Fitts' reciprocal tapping task. *Proc HFES 2009*, pp. 839-843.

76

Other Designs: 1 Factor with 4 Levels



Participant	Test Condition			
	A	B	C	D
1	11	11	21	16
2	18	11	22	15
3	17	10	18	13
4	19	15	21	20
5	13	17	23	10
6	10	15	15	20
7	14	14	15	13
8	13	14	19	18
9	19	18	16	12
10	10	17	21	18
11	10	19	22	13
12	16	14	18	20
13	10	20	17	19
14	10	13	21	18
15	20	17	14	18
16	18	17	17	14
Mean	14.25	15.13	18.75	16.06
SD	3.84	2.94	2.89	3.23

ANOVA Table for Dependent Variable (units)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	15	81.109	5.407				
Test Condition	3	182.172	60.724	4.954	.0047	14.862	.896
Test Condition * Subject	45	551.578	12.257				

77

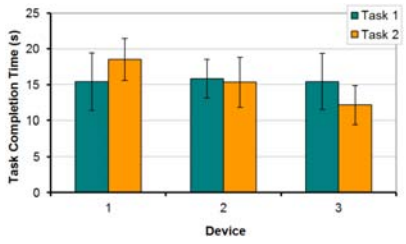
Post Hoc Comparisons

- A significant *F*-test means at least one mean is different from at least one other mean
- Does not reveal which pairs of means are different
- For this, a ***post hoc comparisons*** test is used (aka ***pair-wise comparisons***)
- Example tests
  - Sheffé, Tukey HSD, Fisher LSD, Bonferroni-Dunn

78

Other Designs: 2 Factors

Participant	Device 1		Device 2		Device 3	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
1	11	18	15	13	20	14
2	10	14	17	15	11	13
3	10	23	13	20	20	16
4	18	18	11	12	11	10
5	20	21	19	14	19	8
6	14	21	20	11	17	13
7	14	16	15	20	16	12
8	20	21	18	20	14	12
9	14	15	13	17	16	14
10	20	15	18	10	11	16
11	14	20	15	16	10	9
12	20	20	16	16	20	9
Mean	15.4	18.5	15.8	15.3	15.4	12.2
SD	4.01	2.94	2.69	3.50	3.92	2.69



Main effects (2)

Interaction effect (1)

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	11	134.778	12.253				
Device	2	121.028	60.514	5.865	.0091	11.731	.831
Device * Subject	22	226.972	10.317				
Task	1	.889	.889	.076	.7875	.076	.057
Task * Subject	11	128.111	11.646				
Device * Task	2	121.028	60.514	5.435	.0121	10.869	.798
Device * Task * Subject	22	244.972	11.135				

79

ANOVA Demos

- *StatView* (now sold as JMP, http://jmp.com)
  - Commercial statistics package
  - Input file: `AnovaExample1.svd`
- *Anova2*
  - Java program and its API are available (free download)
  - Input file: `AnovaExample1.txt`
- *PostHoc*
  - Java utility and its API are available (free download)

80

ANOVA Demos (2)

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.080	.564				
Method	1	4.232	4.232	9.796	.0121	9.796	.804
Method * Subject	9	3.888	.432				


```
winterschool>java Anova2 AnovaExample1.txt 10 2 . . -a

ANOVA_table
=============================================================================
Effect                  df        SS          MS          F         p
-----------------------------------------------------------------------------
Participant              9         5.080       0.564
F1                       1         4.232       4.232      9.796     0.0121
F1_x_Par                 9         3.888       0.432
=============================================================================
```

81

Group Participation Results

- Results will be presented in class for the experiment conducted before the break
- The following results are from another run of the same experiment

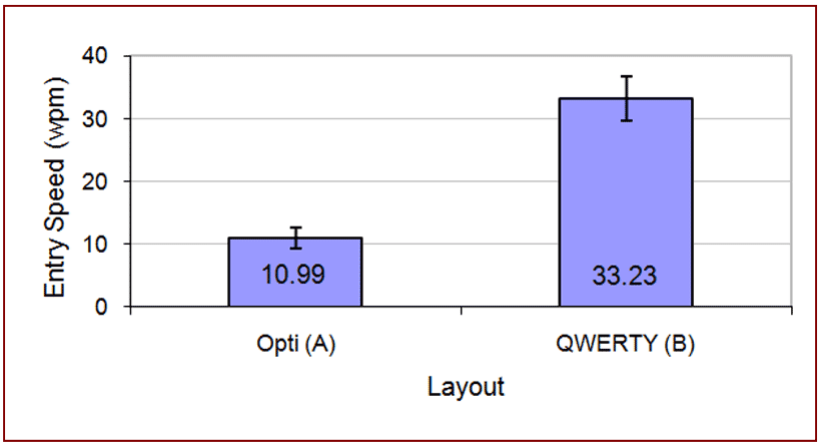
82

Entry Time (seconds)												
Participant	Initials	Opti (A)					QWERTY (B)					Group
		1	2	3	4	5	1	2	3	4	5	
P1	al	92.0	94.0	84.0	68.0	93.0	23.0	19.0	17.0	17.0	15.0	1
P2	ig	65.0	63.0	55.0	49.0	41.0	18.0	15.0	14.0	14.0	13.0	1
P3	ma	54.0	44.0	38.0	38.0	32.0	19.0	17.0	17.0	15.0	19.0	1
P4	kw	65.0	71.0	57.0	61.0	51.0	23.0	19.0	19.0	19.0	18.0	1
P5	ja	40.0	33.0	31.0	29.0	28.0	19.0	17.0	19.0	17.0	16.0	1
P6	ej	66.0	65.0	47.0	52.0	46.0	20.0	17.0	17.0	15.0	14.0	1
P7	ml	50.0	49.0	40.0	36.0	31.0	22.0	18.0	16.0	16.0	14.0	1
P8	pa	68.0	47.0	46.0	35.0	34.0	17.0	13.0	12.0	16.0	12.0	1
P9	ul	86.0	83.0	56.0	46.0	45.0	29.0	19.0	18.0	17.0	15.0	1
P10	em	72.0	67.0	51.0	45.0	49.0	18.0	15.0	13.0	12.0	14.0	1
P11	pl	49.0	48.0	53.0	39.0	39.0	19.0	18.0	17.0	15.0	18.0	1
P12	bc	39.0	43.0	34.0	33.0	32.0	14.0	12.0	13.0	12.0	12.0	1
P13	as	54.0	44.0	41.0	38.0	41.0	17.0	14.0	12.0	13.0	13.0	2
P14	jj	75.0	65.0	55.0	71.0	53.0	21.0	17.0	17.0	19.0	16.0	2
P15	al	83.0	80.0	52.0	67.0	63.0	23.0	22.0	22.0	19.0	18.0	2
P16	sk	60.0	52.0	43.0	39.0	36.0	17.0	19.0	16.0	15.0	15.0	2
P17	jo	84.0	66.0	57.0	40.0	54.0	15.0	13.0	13.0	13.0	12.0	2
P18	hk	74.0	57.0	49.0	45.0	39.0	21.0	20.0	17.0	17.0	16.0	2
P19	mb	58.0	50.0	68.0	51.0	46.0	24.0	18.0	18.0	14.0	14.0	2
P20	jk	64.0	47.0	42.0	41.0	42.0	14.0	14.0	13.0	13.0	12.0	2
P21	ct	60.0	50.0	40.0	39.0	33.0	14.0	12.0	12.0	12.0	11.0	2
P22	hha	62.0	46.0	45.0	40.0	45.0	23.0	18.0	18.0	17.0	16.0	2
P23	ss	37.0	37.0	31.0	31.0	23.0	18.0	14.0	12.0	11.0	11.0	2
P24	ma	49.0	45.0	52.0	43.0	33.0	16.0	13.0	13.0	12.0	12.0	2

83

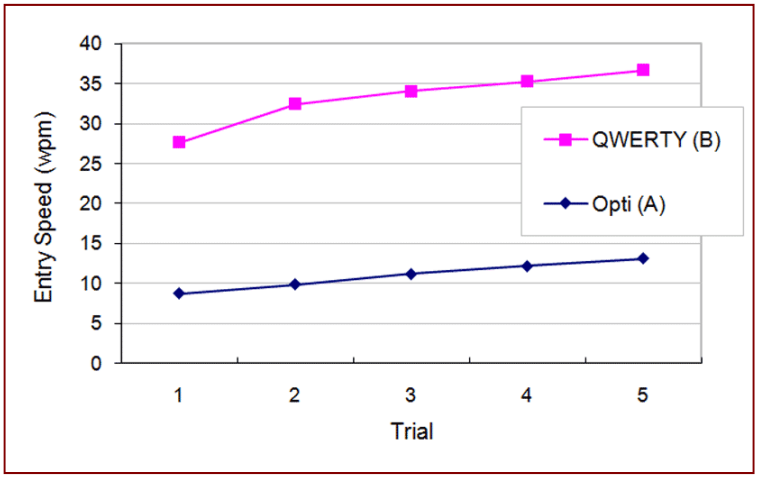
Entry Speed (wpm)												
Participant	Initials	Opti (A)					QWERTY (B)					Group
		1	2	3	4	5	1	2	3	4	5	
P1	al	5.61	5.49	6.14	7.59	5.55	22.43	27.16	30.35	30.35	34.40	1
P2	ig	7.94	8.19	9.38	10.53	12.59	28.67	34.40	36.86	36.86	39.69	1
P3	ma	9.56	11.73	13.58	13.58	16.13	27.16	30.35	30.35	34.40	27.16	1
P4	kw	7.94	7.27	9.05	8.46	10.12	22.43	27.16	27.16	27.16	28.67	1
P5	ja	12.90	15.64	16.65	17.79	18.43	27.16	30.35	27.16	30.35	32.25	1
P6	ej	7.82	7.94	10.98	9.92	11.22	25.80	30.35	30.35	34.40	36.86	1
P7	ml	10.32	10.53	12.90	14.33	16.65	23.45	28.67	32.25	32.25	36.86	1
P8	pa	7.59	10.98	11.22	14.74	15.18	30.35	39.69	43.00	32.25	43.00	1
P9	ul	6.00	6.22	9.21	11.22	11.47	17.79	27.16	28.67	30.35	34.40	1
P10	em	7.17	7.70	10.12	11.47	10.53	28.67	34.40	39.69	43.00	36.86	1
P11	pl	10.53	10.75	9.74	13.23	13.23	27.16	28.67	30.35	34.40	28.67	1
P12	bc	13.23	12.00	15.18	15.64	16.13	36.86	43.00	39.69	43.00	43.00	1
P13	as	9.56	11.73	12.59	13.58	12.59	30.35	36.86	43.00	39.69	39.69	2
P14	jj	6.88	7.94	9.38	7.27	9.74	24.57	30.35	30.35	27.16	32.25	2
P15	al	6.22	6.45	9.92	7.70	8.19	22.43	23.45	23.45	27.16	28.67	2
P16	sk	8.60	9.92	12.00	13.23	14.33	30.35	27.16	32.25	34.40	34.40	2
P17	jo	6.14	7.82	9.05	12.90	9.56	34.40	39.69	39.69	39.69	43.00	2
P18	hk	6.97	9.05	10.53	11.47	13.23	24.57	25.80	30.35	30.35	32.25	2
P19	mb	8.90	10.32	7.59	10.12	11.22	21.50	28.67	28.67	36.86	36.86	2
P20	jk	8.06	10.98	12.29	12.59	12.29	36.86	36.86	39.69	39.69	43.00	2
P21	ct	8.60	10.32	12.90	13.23	15.64	36.86	43.00	43.00	43.00	46.91	2
P22	hha	8.32	11.22	11.47	12.90	11.47	22.43	28.67	28.67	30.35	32.25	2
P23	ss	13.95	13.95	16.65	16.65	22.43	28.67	36.86	43.00	46.91	46.91	2
P24	ma	10.53	11.47	9.92	12.00	15.64	32.25	39.69	39.69	43.00	43.00	2
	Mean	8.72	9.82	11.18	12.17	13.06	27.63	32.43	34.07	35.29	36.71	
	SD	2.27	2.47	2.60	2.77	3.61	5.24	5.74	6.15	5.82	5.91	
					Min	5.49				*Min*	17.79	
					Max	22.43				*Max*	46.91	

84



Note: A *bar chart* is appropriate here because the data along the x-axis are categorical (i.e., nominal scale).

85



Note: A *line chart* is appropriate here because the data along the x-axis are continuous (i.e., ratio scale).

86

ANOVA Table for Entry Speed (wpm)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Group	1	73.737	73.737	.618	.4401	.618	.113
Subject(Group)	22	2624.205	119.282				
Layout	1	29664.381	29664.381	533.785	<.0001	533.785	1.000
Layout * Group	1	80.007	80.007	1.440	.2430	1.440	.199
Layout * Subject(Group)	22	1222.620	55.574				
Trial	4	1298.277	324.569	78.825	<.0001	315.300	1.000
Trial * Group	4	2.688	.672	.163	.9564	.653	.083
Trial * Subject(Group)	88	362.348	4.118				
Layout * Trial	4	172.752	43.188	10.706	<.0001	42.823	1.000
Layout * Trial * Group	4	10.887	2.722	.675	.6113	2.699	.207
Layout * Trial * Subject(Group)	88	354.997	4.034				

- Layout effect is significant ($F_{1,22} = 533.8$, $p < .0001$)
- Trial effect is significant ($F_{4,88} = 78.8$, $p < .0001$)
- Layout by trial interaction effect is significant ($F_{4,88} = 10.7$, $p < .0001$)
- Group effect is not significant ($F_{1,22} = 0.62$, *ns*)

87

Participant	Initials	Sex	Age	English as 1st language	Hours of computer use per day?	Do you regularly use a mobile phone?	Do you send text messages on a mobile phone?	If yes, how many messages per day?
P1	al	Male	43	No	10.0	Yes	Yes	8.0
P2	ig		35		7.0	Yes	n	0.0
P3	ma	female		Yes	8.0	Yes	Yes	5.0
P4	kw	female	33	No	8.0	Yes	Yes	2.5
P5	ja	Male	31	No	10.0	Yes	Yes	20.0
P6	ej	Male	42	Yes	10.0	Yes	Yes	20.0
P7	ml	female	41	No	8.0	Yes	Yes	5.0
P8	pa	Male	39	No	12.0	Yes	Yes	1.0
P9	ul	Male	36	No	10.0	Yes	Yes	3.0
P10	em	Male	45	Yes	8.0	Yes	Yes	5.0
P11	pl	Male	31	No	8.0	Yes	Yes	4.0
P12	bc	female	40	Yes	10.0	Yes	Yes	100.0
P13	as	Male	25	No	8.0	Yes	n	0.0
P14	jj	Male	45	No	6.0	Yes	Yes	5.0
P15	al	Male	51	No	10.0	Yes	Yes	5.0
P16	sk	Male	32	No	8.0	Yes	Yes	10.0
P17	jo	Male	31	No	10.0	Yes	Yes	5.0
P18	hk	female	33	No	10.0	Yes	Yes	20.0
P19	mb	Male	37	No	16.0	Yes	Yes	25.0
P20	jk	female	29	No	8.0	Yes	Yes	1.0
P21	ct	Male	33	Yes	10.0	Yes	Yes	8.0
P22	hha	female	36	No	9.0	n	n	0.0
P23	ss	Male	35	Yes	10.0	Yes	Yes	4.0
P24	ma	female	36	Yes	10.0	Yes	Yes	100.0
Responses		23	23	23	24	24	24	24
Tally		15	839	7	224	23	21	357
Result		65.2%	36.5	30.4%	9.3	95.8%	87.5%	14.9
Units		Male	Years	English	Hours per day	Yes	Yes	Messages per day

88

Topics

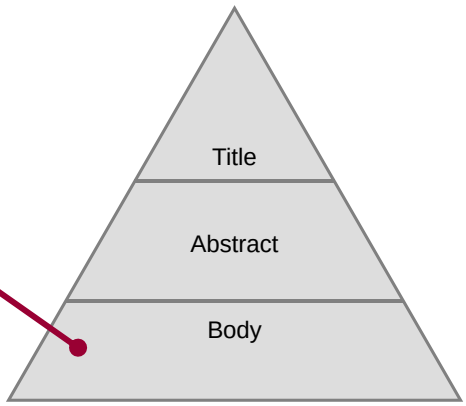
- The what, why, and how
- Group participation in a real experiment
- Observations and measurements
- Research methods (and their properties)
- Experiment terminology
- Experiment design
- ANOVA statistics and experiment results
- Parts of a research paper

89

Research Paper

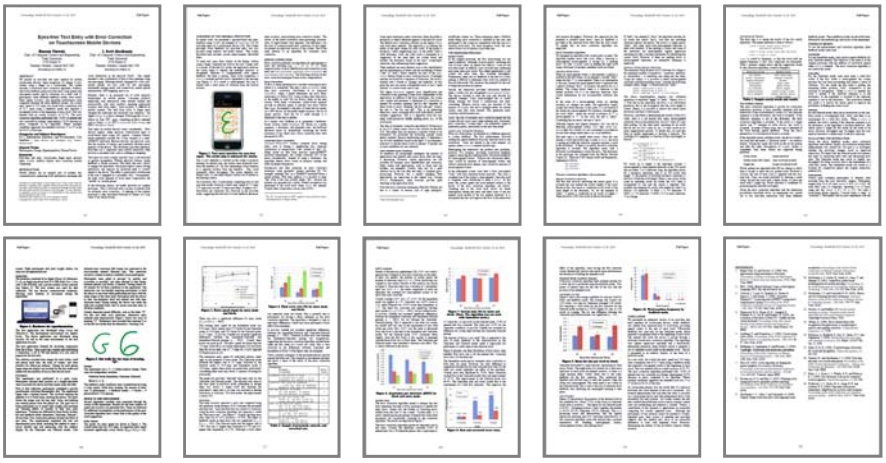
- Research is not finished until the results are published!
- Organization

Main sections…
- Introduction
- Method
  - Participants
  - Apparatus
  - Procedure
  - Design
- Results and Discussion
- Conclusions



90

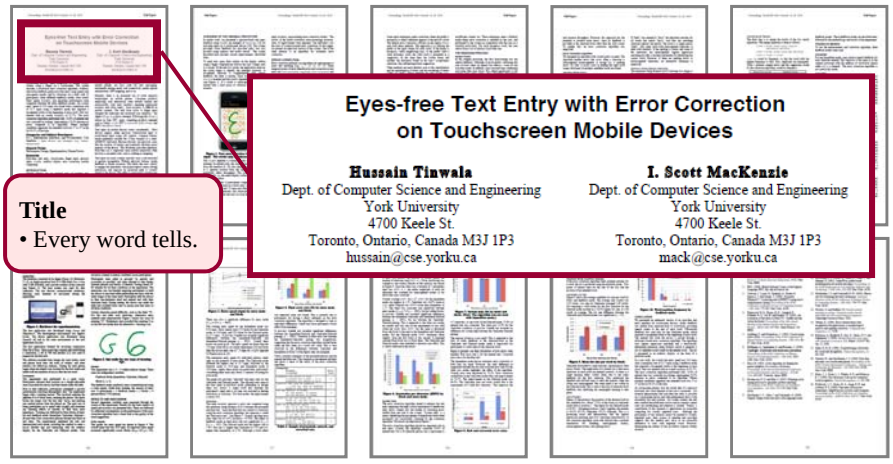
Example Publication†



† Tinwala, H. and MacKenzie, I. S., Eyes-free text entry with error correction on touchscreen mobile devices, *Proceedings of the 6th Nordic Conference on Human-Computer Interaction - NordiCHI 2010*, (New York: ACM, 2010), 511-520.

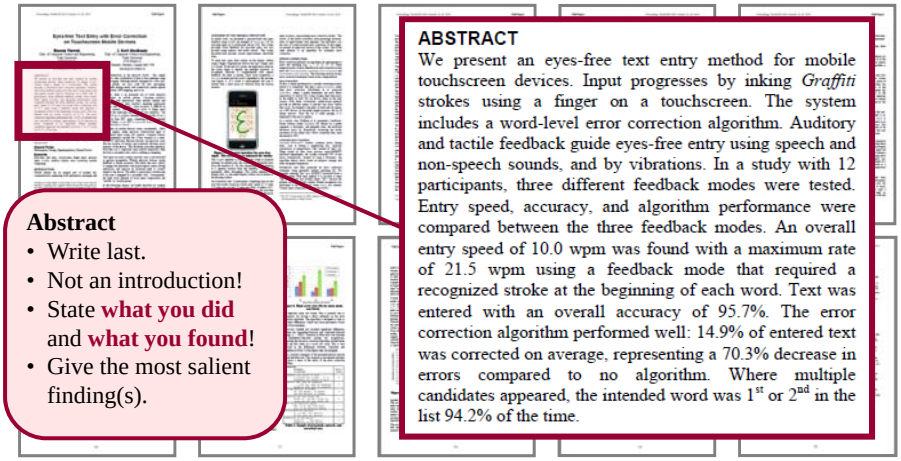
91

Title, Author(s), Affiliation(s)



92

Abstract

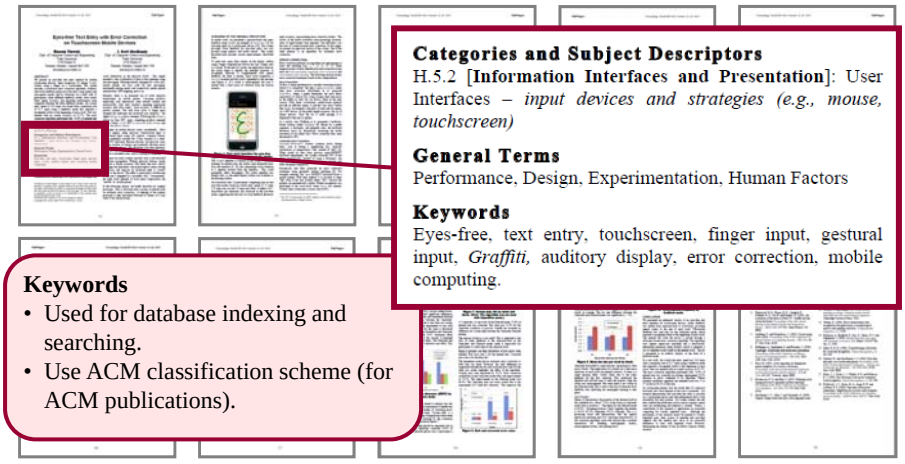


ABSTRACT
We present an eyes-free text entry method for mobile touchscreen devices. Input progresses by inking *Graffiti* strokes using a finger on a touchscreen. The system includes a word-level error correction algorithm. Auditory and tactile feedback guide eyes-free entry using speech and non-speech sounds, and by vibrations. In a study with 12 participants, three different feedback modes were tested. Entry speed, accuracy, and algorithm performance were compared between the three feedback modes. An overall entry speed of 10.0 wpm was found with a maximum rate of 21.5 wpm using a feedback mode that required a recognized stroke at the beginning of each word. Text was entered with an overall accuracy of 95.7%. The error correction algorithm performed well: 14.9% of entered text was corrected on average, representing a 70.3% decrease in errors compared to no algorithm. Where multiple candidates appeared, the intended word was 1st or 2nd in the list 94.2% of the time.

Abstract
- Write last.
- Not an introduction!
- State **what you did** and **what you found**!
- Give the most salient finding(s).

93

Keywords



Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *input devices and strategies (e.g., mouse, touchscreen)*

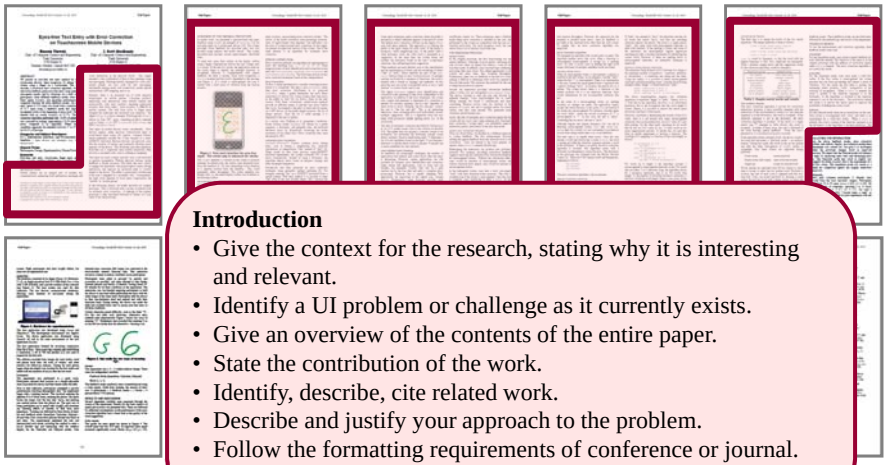
General Terms
Performance, Design, Experimentation, Human Factors

Keywords
Eyes-free, text entry, touchscreen, finger input, gestural input, *Graffiti,* auditory display, error correction, mobile computing.

Keywords
- Used for database indexing and searching.
- Use ACM classification scheme (for ACM publications).

94

Introduction

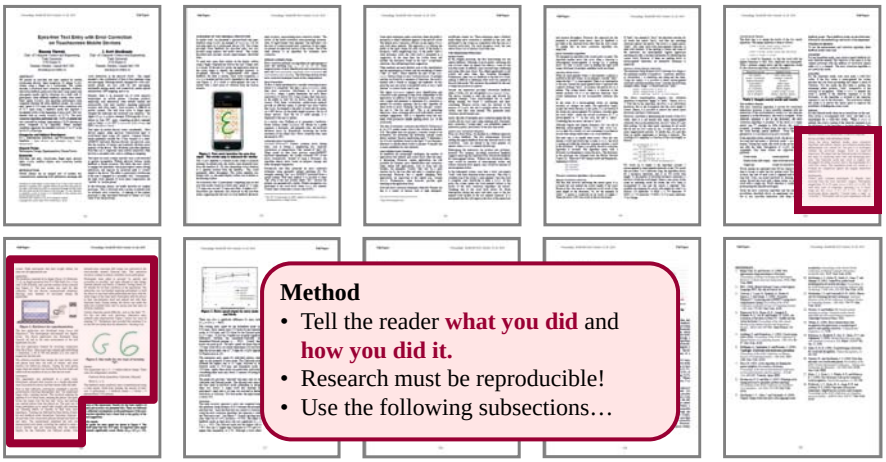


Introduction

- Give the context for the research, stating why it is interesting and relevant.
- Identify a UI problem or challenge as it currently exists.
- Give an overview of the contents of the entire paper.
- State the contribution of the work.
- Identify, describe, cite related work.
- Describe and justify your approach to the problem.
- Follow the formatting requirements of conference or journal.
- **It's your story to tell!**

95

Method

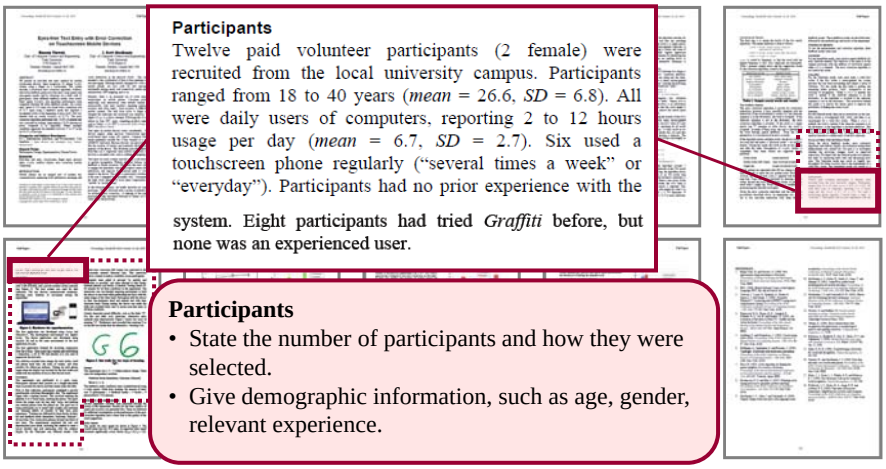


Method

- Tell the reader **what you did** and **how you did it.**
- Research must be reproducible!
- Use the following subsections…

96

Method - Participants

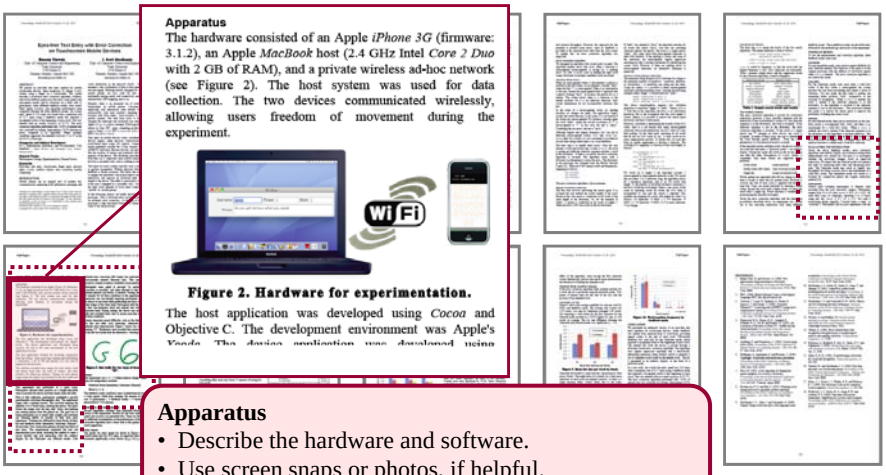


Participants
Twelve paid volunteer participants (2 female) were recruited from the local university campus. Participants ranged from 18 to 40 years (*mean* = 26.6, *SD* = 6.8). All were daily users of computers, reporting 2 to 12 hours usage per day (*mean* = 6.7, *SD* = 2.7). Six used a touchscreen phone regularly ("several times a week" or "everyday"). Participants had no prior experience with the system. Eight participants had tried *Graffiti* before, but none was an experienced user.

Participants
- State the number of participants and how they were selected.
- Give demographic information, such as age, gender, relevant experience.

97

Method - Apparatus



Apparatus
The hardware consisted of an Apple *iPhone 3G* (firmware: 3.1.2), an Apple *MacBook* host (2.4 GHz Intel *Core 2 Duo* with 2 GB of RAM), and a private wireless ad-hoc network (see Figure 2). The host system was used for data collection. The two devices communicated wirelessly, allowing users freedom of movement during the experiment.

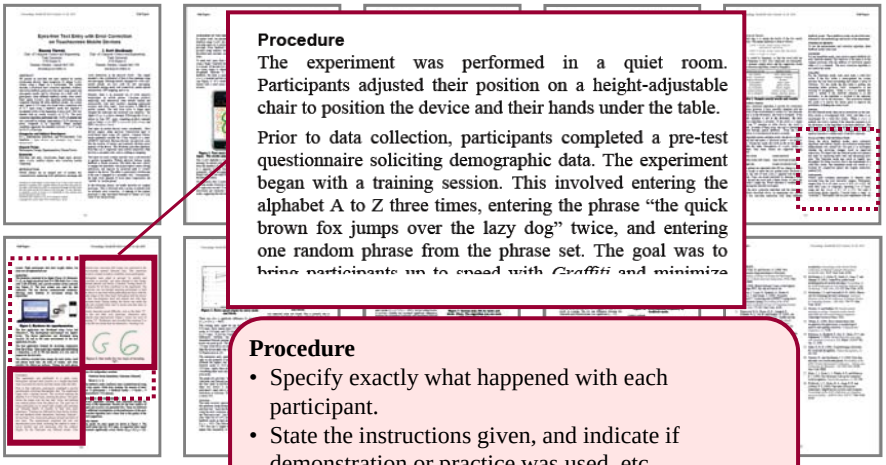
Figure 2. Hardware for experimentation.

The host application was developed using *Cocoa* and Objective C. The development environment was Apple's

Apparatus
- Describe the hardware and software.
- Use screen snaps or photos, if helpful.

98

Method - Procedure



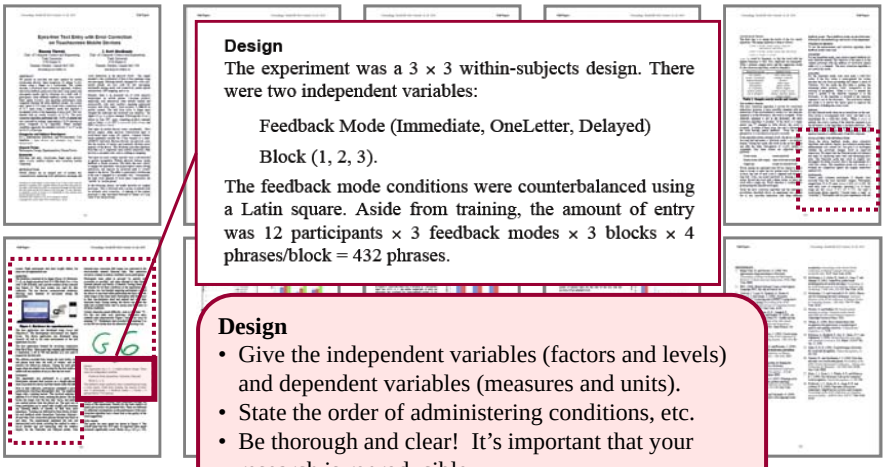
Procedure
The experiment was performed in a quiet room. Participants adjusted their position on a height-adjustable chair to position the device and their hands under the table.

Prior to data collection, participants completed a pre-test questionnaire soliciting demographic data. The experiment began with a training session. This involved entering the alphabet A to Z three times, entering the phrase "the quick brown fox jumps over the lazy dog" twice, and entering one random phrase from the phrase set. The goal was to bring participants up to speed with *Graffiti* and minimize

Procedure
- Specify exactly what happened with each participant.
- State the instructions given, and indicate if demonstration or practice was used, etc.

99

Method - Design



Design
The experiment was a 3 × 3 within-subjects design. There were two independent variables:

Feedback Mode (Immediate, OneLetter, Delayed)

Block (1, 2, 3).

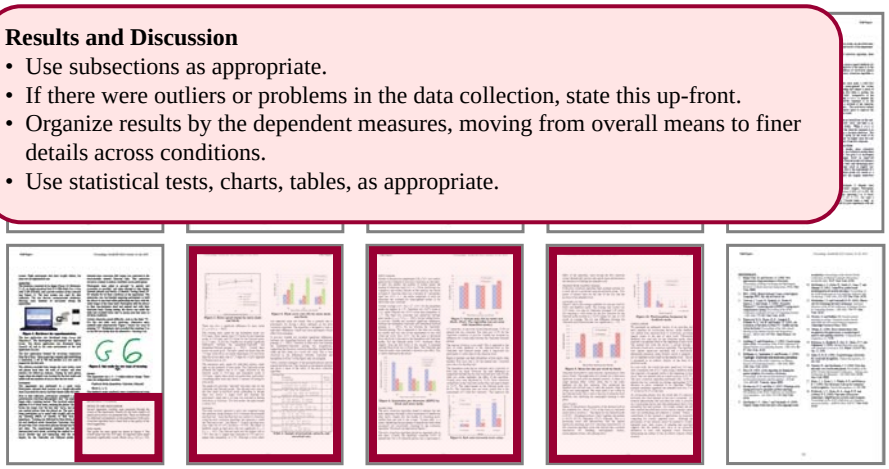
The feedback mode conditions were counterbalanced using a Latin square. Aside from training, the amount of entry was 12 participants × 3 feedback modes × 3 blocks × 4 phrases/block = 432 phrases.

Design
- Give the independent variables (factors and levels) and dependent variables (measures and units).
- State the order of administering conditions, etc.
- Be thorough and clear! It's important that your research is reproducible.

100

Results and Discussion

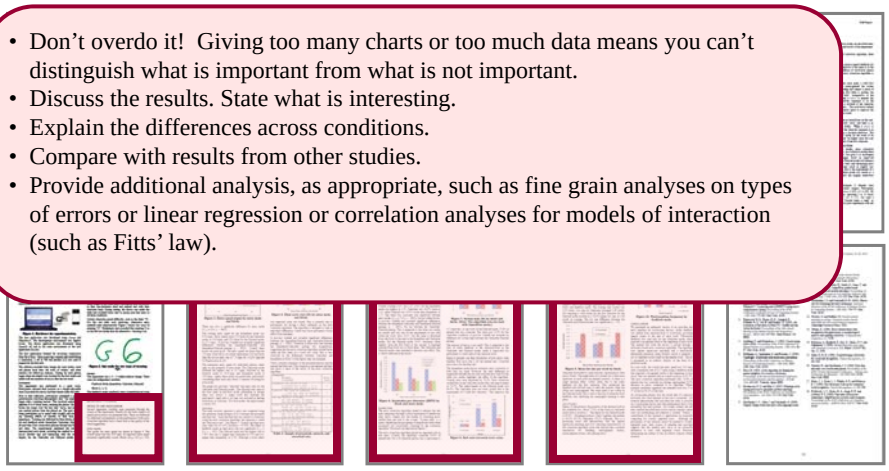
Results and Discussion
- Use subsections as appropriate.
- If there were outliers or problems in the data collection, state this up-front.
- Organize results by the dependent measures, moving from overall means to finer details across conditions.
- Use statistical tests, charts, tables, as appropriate.



101

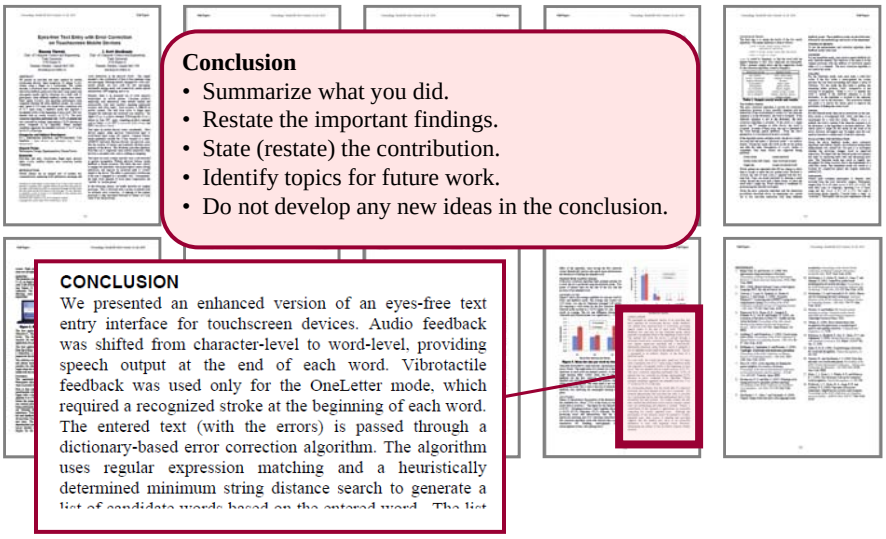
Results and Discussion (2)

- Don't overdo it! Giving too many charts or too much data means you can't distinguish what is important from what is not important.
- Discuss the results. State what is interesting.
- Explain the differences across conditions.
- Compare with results from other studies.
- Provide additional analysis, as appropriate, such as fine grain analyses on types of errors or linear regression or correlation analyses for models of interaction (such as Fitts' law).



102

Conclusion



Conclusion
- Summarize what you did.
- Restate the important findings.
- State (restate) the contribution.
- Identify topics for future work.
- Do not develop any new ideas in the conclusion.

CONCLUSION

We presented an enhanced version of an eyes-free text entry interface for touchscreen devices. Audio feedback was shifted from character-level to word-level, providing speech output at the end of each word. Vibrotactile feedback was used only for the OneLetter mode, which required a recognized stroke at the beginning of each word. The entered text (with the errors) is passed through a dictionary-based error correction algorithm. The algorithm uses regular expression matching and a heuristically determined minimum string distance search to generate a

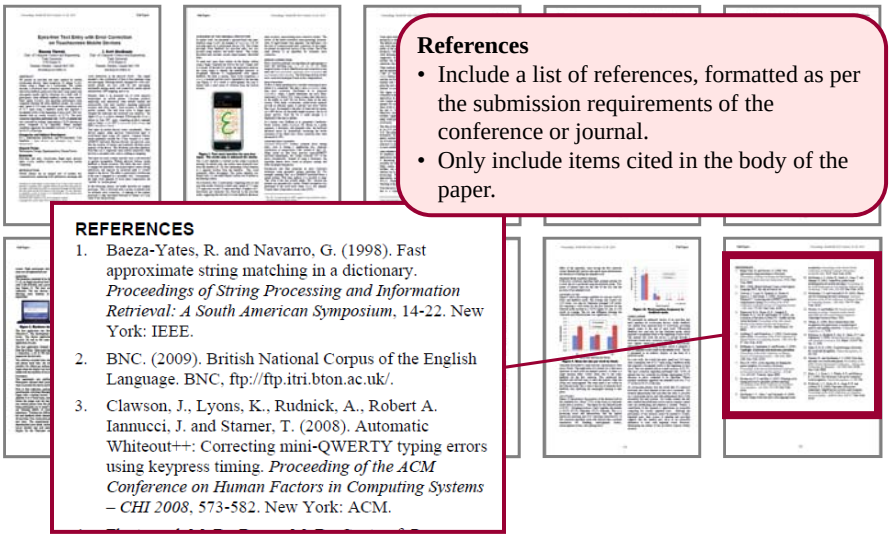
103

Acknowledgment

Acknowledgment
- Optional.
- Thank people who helped.
- Thank funding agencies.

104

References



References
- Include a list of references, formatted as per the submission requirements of the conference or journal.
- Only include items cited in the body of the paper.

REFERENCES

1. Baeza-Yates, R. and Navarro, G. (1998). Fast approximate string matching in a dictionary. *Proceedings of String Processing and Information Retrieval: A South American Symposium*, 14-22. New York: IEEE.
2. BNC. (2009). British National Corpus of the English Language. BNC, ftp://ftp.itri.bton.ac.uk/.
3. Clawson, J., Lyons, K., Rudnick, A., Robert A. Iannucci, J. and Starner, T. (2008). Automatic Whiteout++: Correcting mini-QWERTY typing errors using keypress timing. *Proceeding of the ACM Conference on Human Factors in Computing Systems – CHI 2008*, 573-582. New York: ACM.

105

Summary

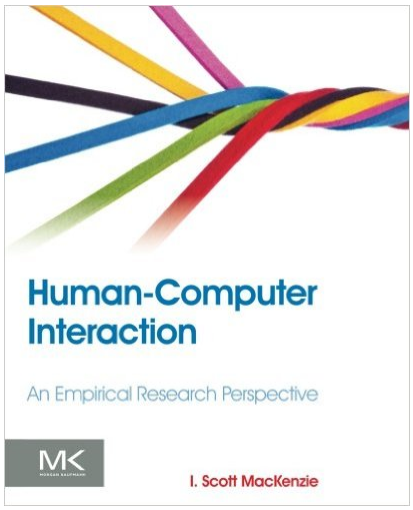
- The what, why, and how of empirical research
- Group participation in a real experiment
- Observations and measurements
- Research methods (and their properties)
- Experiment terminology
- Experiment design
- ANOVA statistics and experiment results
- Parts of a research paper

Thank you

http://www.yorku.ca/mack/UCLAN

106

For the complete story, see Scott's book:



http://www.yorku.ca/mack/HCIbook

107