Empirical Research Methods for Human-Computer Interaction

I. Scott MacKenzie
Steven J. Castellucci

York University
Toronto, Canada

CHI 2016 San Jose, CA, USA May 7-12

Presenters

Scott MacKenzie's MacKenzie's research is in HCI with an emphasis on human performance measurement and modeling, experimental methods and evaluation, interaction devices and techniques, alphanumeric entry, language modeling, and mobile computing. He has more than 160 HCI publications (including more than 40 from the SIGCHI conference and 2 HCI books) and has given numerous invited talks over the past 20 years. Since 1999, he has been Associate Professor of Electrical Engineering and Computer Science at York University, Canada.

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

Steven Castellucci is an Assistant Professor of Electrical Engineering and Computer Science at York University, Canada. His research interests include mobile text entry, and remote pointing techniques. In addition to having SIGCHI publications, he teaches first-year Computer Science and Engineering courses, and senior-level HCI university courses.

Home page: http://www.eecs.yorku.ca/~stevenc/

Agenda

Session One

- Opening remarks and introduction to empirical research methods (what, why, how)
- Hands-on experiment with group participation
- Observations and measurements (how to gather empirical data)
- Research questions (how to formulate good, testable research questions)
- Experiment terminology (how to describe aspects of a user study)

Session Two

- Experiment design (how to plan and conduct a user study)
- Analysis of variance statistics explained with results from the hands-on experiment in Session One
- Parts of a research paper (required sections, required content, tips on producing a successful research paper)

Topics

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

What is Empirical Research?

- Empirical Research is…
  - Experimentation to discover and interpret facts, revise theories or laws
  - Capable of being verified or disproved by observation or experiment
- In HCI, we focus on phenomena surrounding humans interacting with computers

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)

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

CHI 2016 San Jose, CA, USA May 7-12

Topics

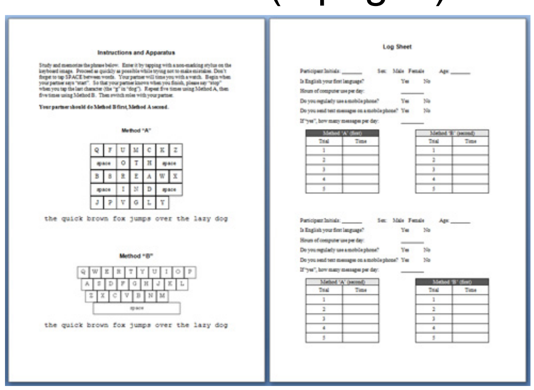
- The what, why, and how of empirical research
- Group participation in a real experiment
- Observations and measurements
- Research questions
- Experiment terminology
- Experiment design
- ANOVA statistics and experiment results
- Parts of a research paper

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

CHI 2016 San Jose, CA, USA May 7-12

Handout (2 pages)



Full-size copies of the handout pages will be distributed during the course. The pages are also contained in an appendix to this package.

CHI 2016 San Jose, CA, USA May 7-12

Do the Experiment

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

I. Scott MacKenzie and Steven J. Castellucci 11

CHI 2016 San Jose, CA, USA May 7-12

Topics

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

I. Scott MacKenzie and Steven J. Castellucci 12

Observations and Measurements

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

CHI 2016 San Jose, CA, USA May 7-12

Scales of Measurement[1]

- Nominal
- Ordinal
- Interval
- Ratio

crude

sophisticated

[1] Stevens S.S. (1946). "On the Theory of Scales of Measurement". *Science* 103 (2684), pp. 677-680.

Nominal Data

- Nominal data (aka categorical data) are arbitrary codes assigned to attributes; e.g.,
  - M = male, F = female
  - 1 = mouse, 2 = touchpad, 3 = pointing stick
- Obviously, the statistical mean cannot be computed on nominal data
- Usually it is the count that is important
  - "Are females or males more likely to…"
  - "Do left or right handers have more difficulty with…"
  - Note: The count itself is a ratio-scale measurement

Nominal Data Example In HCI

- Observe students "on the move" on university campus
- Code and count students by…
  - Gender (male, female)
  - Mobile phone usage (not using, using)

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%		

Real Data !

Ordinal Data

- Ordinal data associate order or rank to an attribute
- The attribute is any characteristic or circumstance of interest; e.g.,
  - Users try three different GPS systems for a period of time, then rank them: 1^{st}, 2^{nd}, 3^{rd} choice
- More sophisticated than nominal data
  - Comparisons of “greater than” or “less than” possible

Ordinal Data Example in HCI

How many text messages do you send each day?

1. Less than 10 per day
2. 10-50 per day
3. 51-99 per day
4. 100-200 per day
5. More than 200 per day

Interval Data

- Equal distances between adjacent values
- But, no absolute zero
- Classic example: temperature (°F, °C)
- Statistical mean possible
  - E.g., the mean midday temperature during July
- Ratios not possible
  - Cannot say 10 °C is twice 5 °C

CHI 2016 San Jose, CA, USA May 7-12

Interval Data Example in HCI

- Questionnaires often solicit a level of agreement to a statement
- Responses on a Likert scale
- Likert scale characteristics:
  1. Statement soliciting level of agreement
  2. Responses are symmetric about a neutral middle value
  3. Gradations between responses are equal (more-or-less)
- Assuming "equal gradations", the statistical mean is valid (and related statistical tests are possible)

CHI 2016 San Jose, CA, USA May 7-12

Interval Data Example in HCI (2)

Please indicate your level of agreement with the following statements.

	Strongly disagree	Mildly disagree	Neutral	Mildly agree	Strongly agree
It is safe to talk on a mobile phone while driving.	1	2	3	4	5
It is safe to compose a text message on a mobile phone while driving.	1	2	3	4	5
It is safe to read a text message on a mobile phone while driving.	1	2	3	4	5

I. Scott MacKenzie and Steven J. Castellucci 21

CHI 2016 San Jose, CA, USA May 7-12

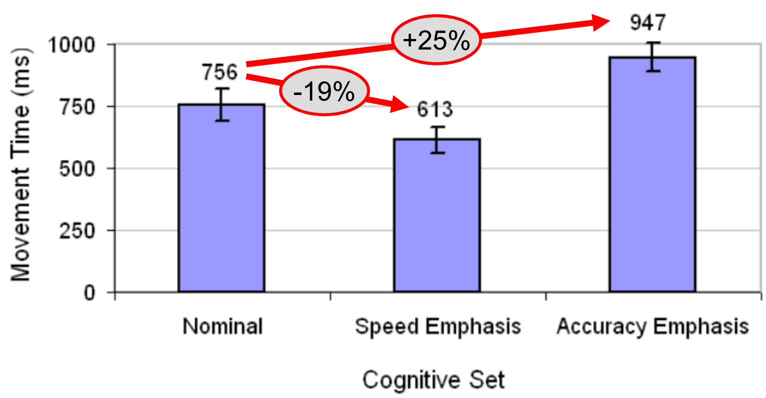
Ratio Data

- Most sophisticated of the four scales of measurement
- Preferred scale of measurement
- Absolute zero, therefore many calculations possible
- Summaries and comparisons are strengthened
- A "count" is a ratio-scale measurement
  - E.g., "time" (the number of seconds to complete a task)
- Enhance counts by adding further ratios where possible
  - Facilitates comparisons
  - Example – a 10-word phrase was entered in 30 seconds
    - Bad: t = 30 seconds (0.5 minutes)
    - Good: Entry rate = 10 / 0.5 = 20 wpm (words-per-minute)

I. Scott MacKenzie and Steven J. Castellucci 22

CHI 2016 San Jose, CA, USA May 7-12

Ratio Data Example in HCI



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

I. Scott MacKenzie and Steven J. Castellucci 23

CHI 2016 San Jose, CA, USA May 7-12

Topics

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

I. Scott MacKenzie and Steven J. Castellucci 24

Research Questions

- Consider the following questions:
  - Is it viable?
  - Is it better than current practice?
  - Which design alternative is best?
  - 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

Testable Research Questions

- 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, and you think it is better than the existing Qwerty soft keyboard (QSK)
  - You decide to undertake a program of empirical enquiry to evaluate your invention
  - What are your research questions?

Research Questions Revisited

- Very weak

    Is the new technique any good?

- Weak

    Is the new technique better than QSK?

- Better

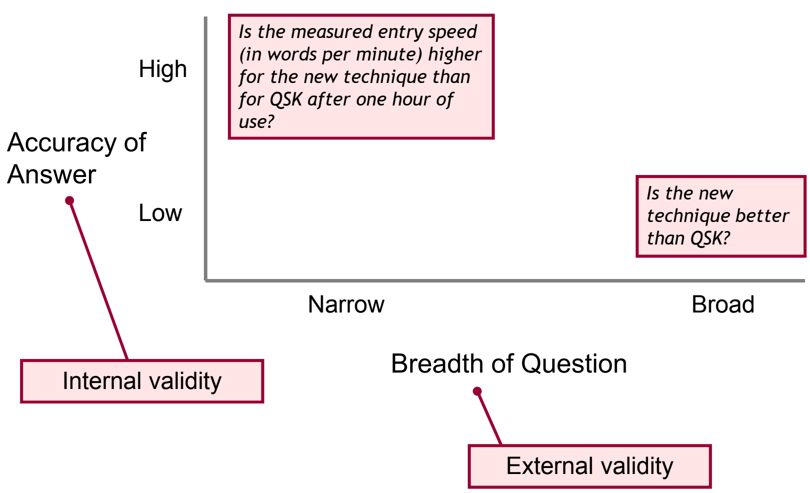
    Is the new technique faster than QSK?

- Better still

    Is the measured entry speed (in words per minute) higher for the new technique than for QSK after one hour of use?

CHI 2016 San Jose, CA, USA May 7-12

A Tradeoff



CHI 2016 San Jose, CA, USA May 7-12

Internal Validity

- Definition:
  - The extent to which the effects observed are due to the ***test conditions*** (e.g., QSK vs. new)
- This means…
  - Differences (in the means) are due to ***inherent properties*** of the test conditions
  - Variances are due to ***participant pre-dispositions***, are controlled, or exist equally across the test conditions

I. Scott MacKenzie and Steven J. Castellucci 29

CHI 2016 San Jose, CA, USA May 7-12

External Validity

- Definition:
  - The extent to which results are generalizable to other ***people*** and other ***situations***
- This 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

I. Scott MacKenzie and Steven J. Castellucci 30

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; so…
  - Use a large display or projection screen (not a desktop monitor)
  - 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?

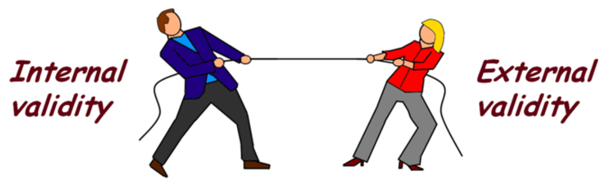
CHI 2016 San Jose, CA, USA May 7-12

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; so …
- Test procedure should probably have participants…
  - Enter representative samples of text (e.g., phrases containing letters, numbers, punctuation, etc.)
  - Edit and correct mistakes as they normally would
- But… is internal validity compromised?

CHI 2016 San Jose, CA, USA May 7-12

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

I. Scott MacKenzie and Steven J. Castellucci 33

CHI 2016 San Jose, CA, USA May 7-12

Topics

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

I. Scott MacKenzie and Steven J. Castellucci 34

Experiment Terminology (Part 1)

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

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*** (e.g., "all ***participants*** exhibited a high error rate…")
- General discussion on the problem or conclusions may use other terms (e.g., "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.

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.

Independent Variable

- An ***independent variable*** is 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 include interface, 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 a name for both the ***factor*** (***independent variable***) and its ***levels*** (***test conditions***)
- Examples

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

Dependent Variable

- A ***dependent variable*** is any measurable aspect of the interaction involving an independent variable
- Examples include 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 (e.g., “text entry speed” is a dependent variable with units “words per minute”)
- Make sure you clearly define all dependent variables
- Research must be reproducible!

Control Variable

- A ***control variable*** is a circumstance (not under investigation) that is kept constant to test the effect of an independent variable
- More control means the experiment is less generalizable (i.e., less applicable to other people and other situations)
- 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)

Confounding Variable

- A ***confounding variable*** is a circumstance that varies systematically with an independent variable
- Should be controlled or randomized to avoid misleading results
- Consider a study comparing the target selection performance of a mouse and a gamepad where all participants are mouse experts, but gamepad novices
  - Mouse performance will likely be higher, but…
  - "Prior experience" is a confounding variable
  - No reliable conclusions can be made

Topics

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

CHI 2016 San Jose, CA, USA May 7-12

Experiment Design

- ***Experiment design*** is 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…

Experiment Terminology (Part 2)

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

Within-subjects, Between-subjects (1)

- 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

Within-subjects, Between-subjects (2)

- 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!)
- Within-subjects disadvantage
  - Order effects (i.e., interference between conditions)
- Between-subjects advantage
  - No order effects (i.e., no interference between conditions)
- Between-subjects disadvantage
  - 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)

Within-subjects, Between-subjects (3)

- 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*)

Counterbalancing

- For within-subjects designs, participants may benefit from the first condition and consequently perform better on the second condition – we don't want this!
- 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)
- ***Group***, then, is a between subjects factor
  - Was there an effect for group? Hopefully not!

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 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)

CHI 2016 San Jose, CA, USA May 7-12

Topics

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

Answering Research Questions

- We want to know if the measured performance on a 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?

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 statistically significant (or is it due to chance)?
   - Statistics can help!
   - The statistical tool is the analysis of variance (ANOVA)

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.

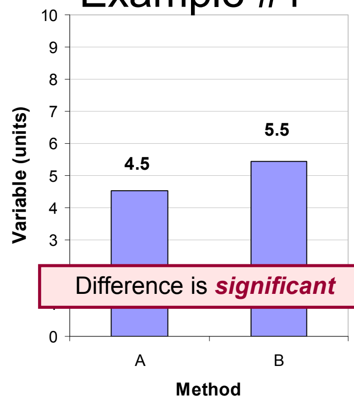
- Assumption of "no difference"
- Research seeks to reject the null hypothesis
- Please bear in mind, with experimental research…
  - We gather and test evidence
  - We do not prove things

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?

CHI 2016 San Jose, CA, USA May 7-12

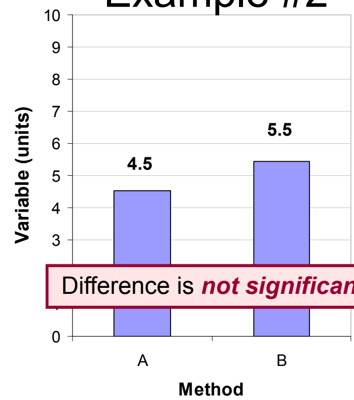
Example #1



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***

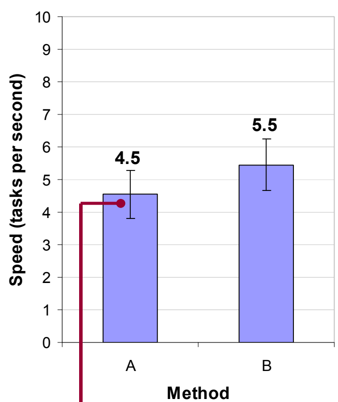
"Not significant" implies that the difference observed is likely due to chance.

File: AnovaDemo.xls

I. Scott MacKenzie and Steven J. Castellucci 57

CHI 2016 San Jose, CA, USA May 7-12

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

I. Scott MacKenzie and Steven J. Castellucci 58

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

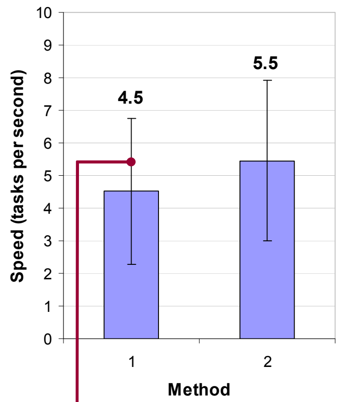
How to Report an *F*-statistic

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 significant figures for *F* statistic
  - No zero before the decimal point in the *p* statistic (except in Europe)

2016 San Jose, CA, USA May 7-12

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

I. Scott MacKenzie and Steven J. Castellucci 61

2016 San Jose, CA, USA May 7-12

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…

$F_{1,9} = 0.449$, ns

Two ways of reporting non-significant effects:

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

I. Scott MacKenzie and Steven J. Castellucci 62

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

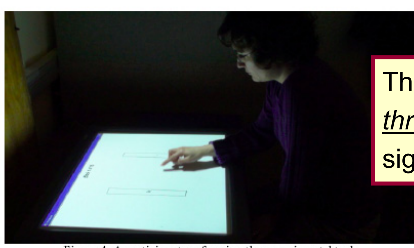


Figure 4. A participant performing the experimental task

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 41.1% 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

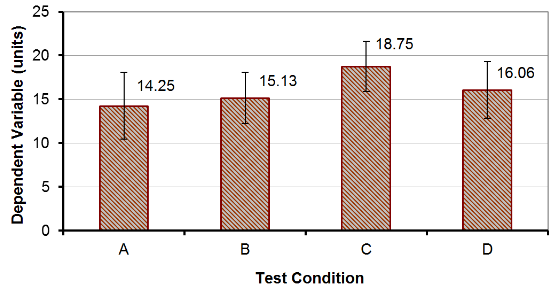
The effect of <u>*input technique*</u> on <u>*throughput*</u> was statistically significant ($F_{1,11}$ = 35.51, p < .0001).

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. Santa Monica, CA: HFES.

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

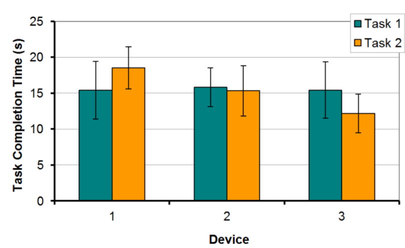
ANOVA Table for Dependent Variable (units)

	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				

I. Scott MacKenzie and Steven J. Castellucci 65

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				

I. Scott MacKenzie and Steven J. Castellucci 66

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

I. Scott MacKenzie and Steven J. Castellucci 67

CHI 2016 San Jose, CA, USA May 7-12

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)

I. Scott MacKenzie and Steven J. Castellucci 68

ANOVA Demos (2)

StatView

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				

Anova2

```
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
===========================================================================
```


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

Entry Time (seconds)

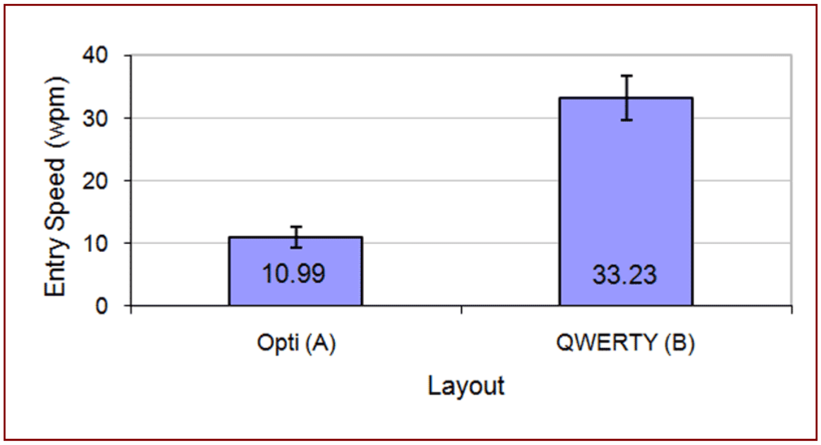
Participant	Initials	Opti (A)					QWERTY (B)					Group
		1	2	3	4	5	1	2	3	4	5	
P1	al	92.0	94.0	84.0	68.0	93.0	23.0	19.0	17.0	17.0	15.0	1
P2	ig	65.0	63.0	55.0	49.0	41.0	18.0	15.0	14.0	14.0	13.0	1
P3	ma	54.0	44.0	38.0	38.0	32.0	19.0	17.0	17.0	15.0	19.0	1
P4	kw	65.0	71.0	57.0	61.0	51.0	23.0	19.0	19.0	19.0	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

I. Scott MacKenzie and Steven J. Castellucci 71

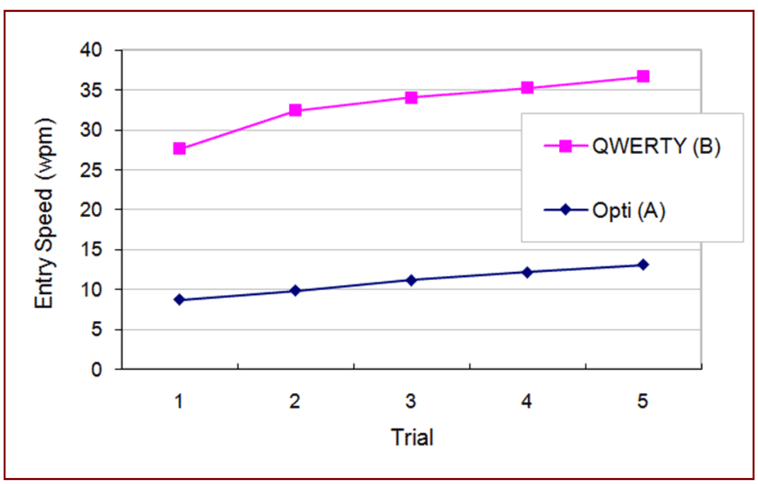
Entry Speed (wpm)

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	

I. Scott MacKenzie and Steven J. Castellucci 72



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



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

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*)

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

San Jose, CA, USA May 7-12

Topics

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

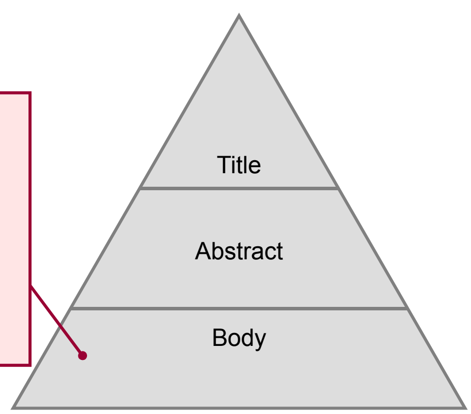
I. Scott MacKenzie and Steven J. Castellucci 77

San Jose, CA, USA May 7-12

Research Paper

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

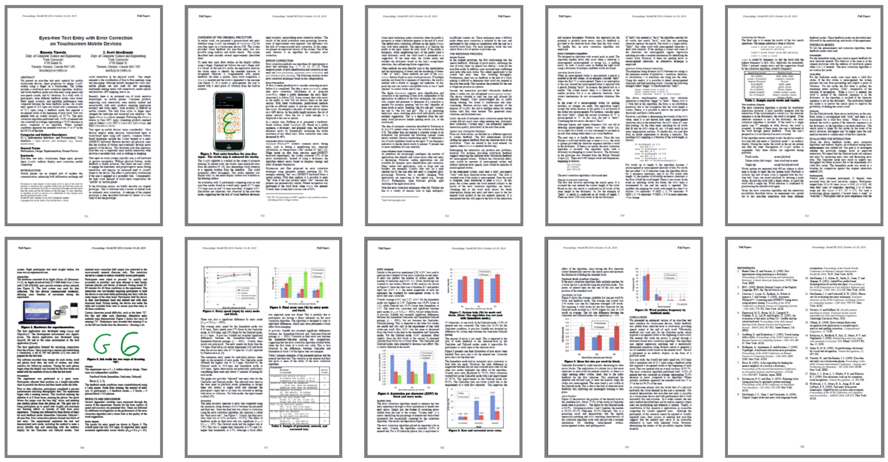
Main sections…
- Introduction
- Method
    - Participants
    - Apparatus
    - Procedure
    - Design
- Results and Discussion
- Conclusions



I. Scott MacKenzie and Steven J. Castellucci 78

CHI 2016 San Jose, CA, USA May 7-12

Example Publication†

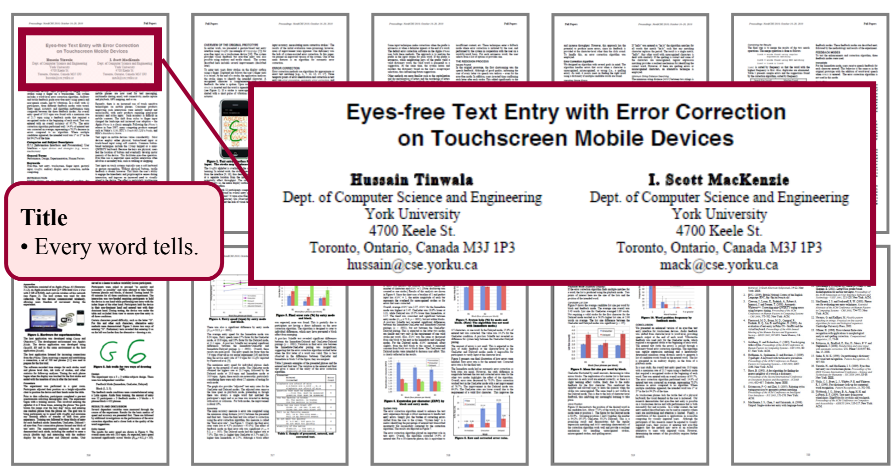


† Tinwala, H. and MacKenzie, I. S., Eyes-free text entry with error correction on touchscreen mobile devices, *Proceedings of the 6th Nordic Conference on Human-Computer Interaction - NordiCHI 2010*, (New York: ACM, 2010), 511-520.

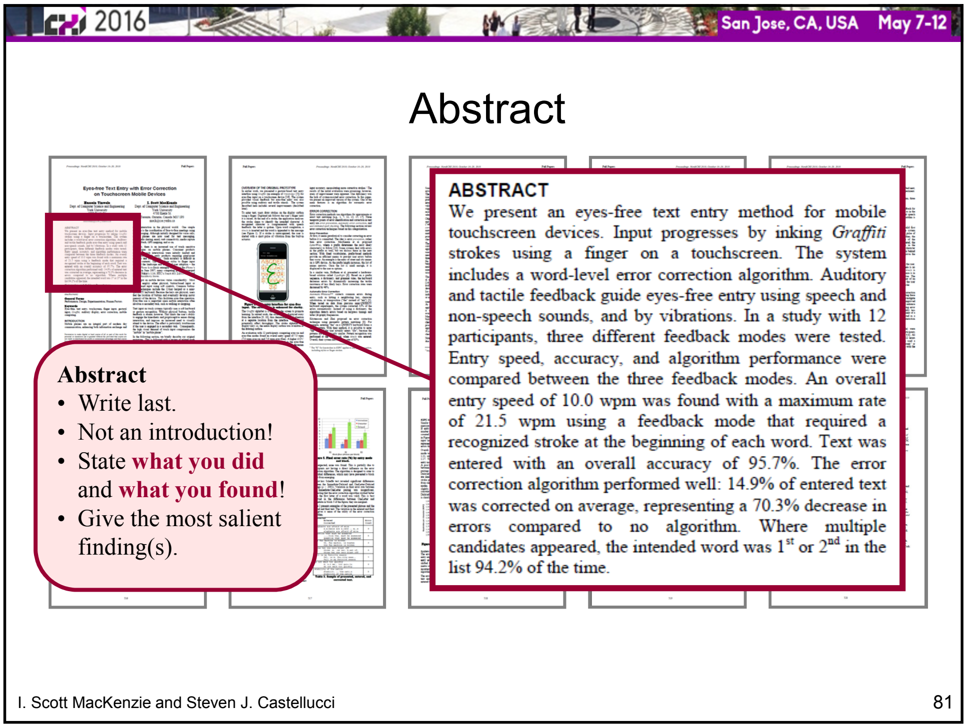
I. Scott MacKenzie and Steven J. Castellucci 79

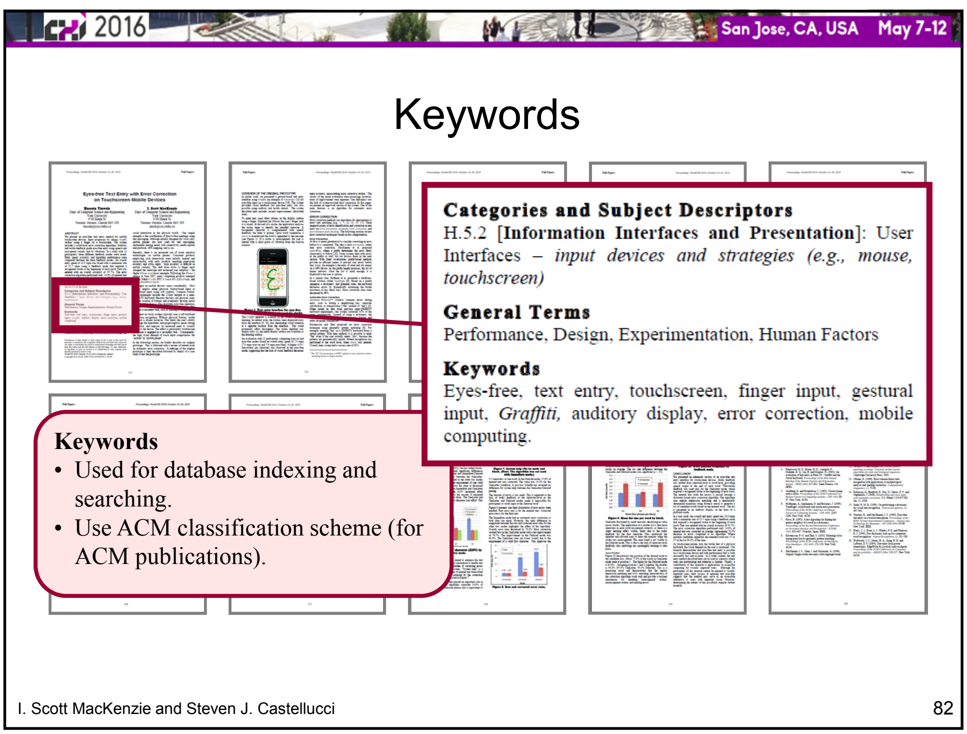
CHI 2016 San Jose, CA, USA May 7-12

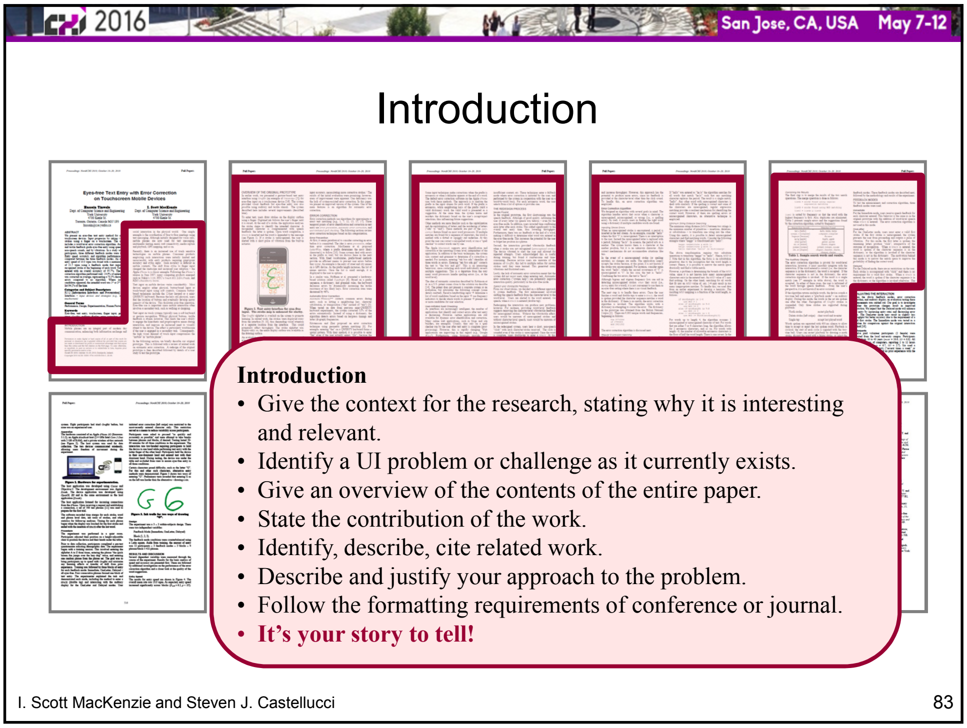
Title, Author(s), Affiliation(s)

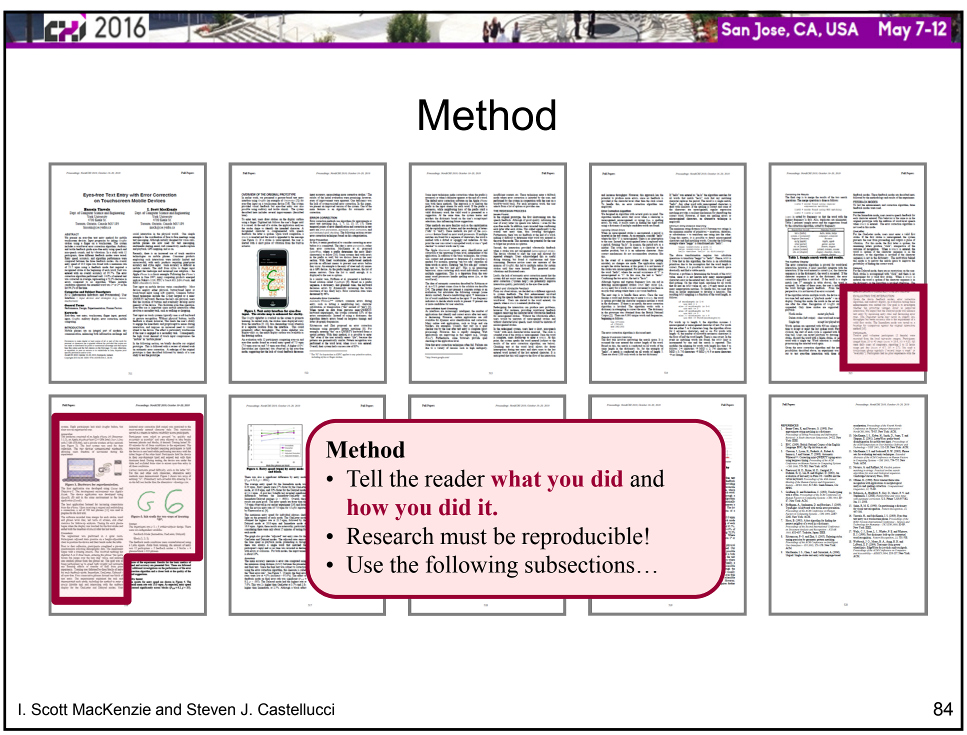


I. Scott MacKenzie and Steven J. Castellucci 80


CHI 2016
San Jose, CA, USA May 7-12
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).
I. Scott MacKenzie and Steven J. Castellucci
81


CHI 2016
San Jose, CA, USA May 7-12
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).
I. Scott MacKenzie and Steven J. Castellucci
82


2016
San Jose, CA, USA May 7-12
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!
I. Scott MacKenzie and Steven J. Castellucci
83


2016
San Jose, CA, USA May 7-12
Method
Method
• Tell the reader what you did and how you did it.
• Research must be reproducible!
• Use the following subsections…
I. Scott MacKenzie and Steven J. Castellucci
84

CHI 2016 San Jose, CA, USA May 7-12

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.

I. Scott MacKenzie and Steven J. Castellucci 85

CHI 2016 San Jose, CA, USA May 7-12

Method - Apparatus

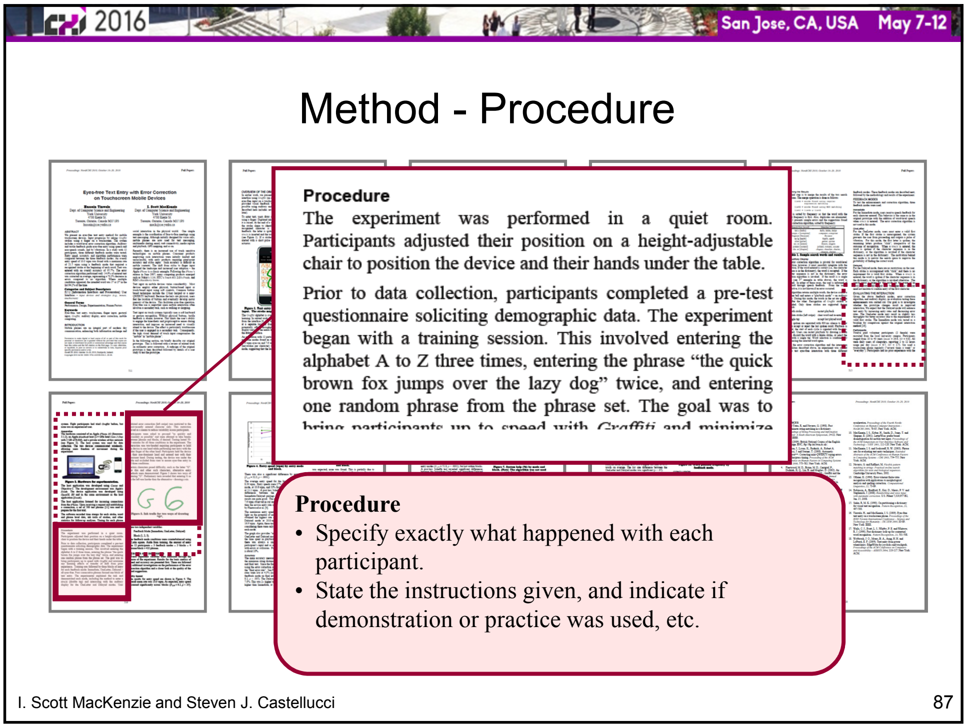
Apparatus
The hardware consisted of an Apple *iPhone 3G* (firmware: 3.1.2), an Apple *MacBook* host (2.4 GHz Intel *Core 2 Duo* with 2 GB of RAM), and a private wireless ad-hoc network (see Figure 2). The host system was used for data collection. The two devices communicated wirelessly, allowing users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

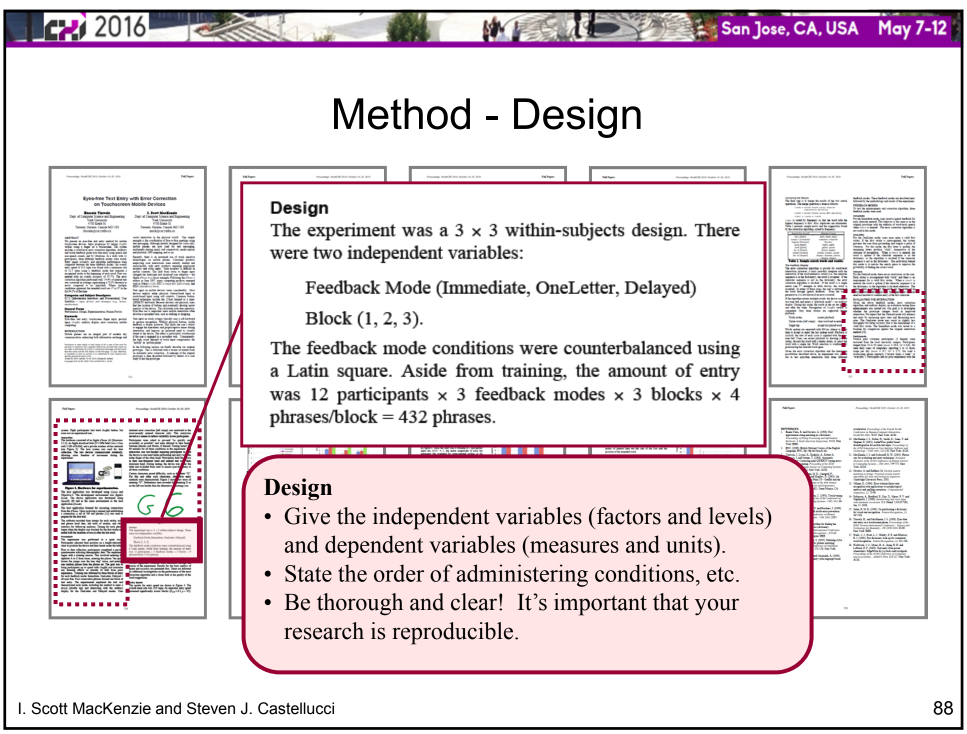
Figure 2. Hardware for experimentation.

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.

I. Scott MacKenzie and Steven J. Castellucci 86


CHI 2016
San Jose, CA, USA May 7-12
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
Procedure
• Specify exactly what happened with each participant.
• State the instructions given, and indicate if demonstration or practice was used, etc.
I. Scott MacKenzie and Steven J. Castellucci
87


CHI 2016
San Jose, CA, USA May 7-12
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.
I. Scott MacKenzie and Steven J. Castellucci
88

CHI 2016 San Jose, CA, USA May 7-12

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.

I. Scott MacKenzie and Steven J. Castellucci 89

CHI 2016 San Jose, CA, USA May 7-12

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).

I. Scott MacKenzie and Steven J. Castellucci 90

CHI 2016 San Jose, CA, USA May 7-12

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

I. Scott MacKenzie and Steven J. Castellucci 91

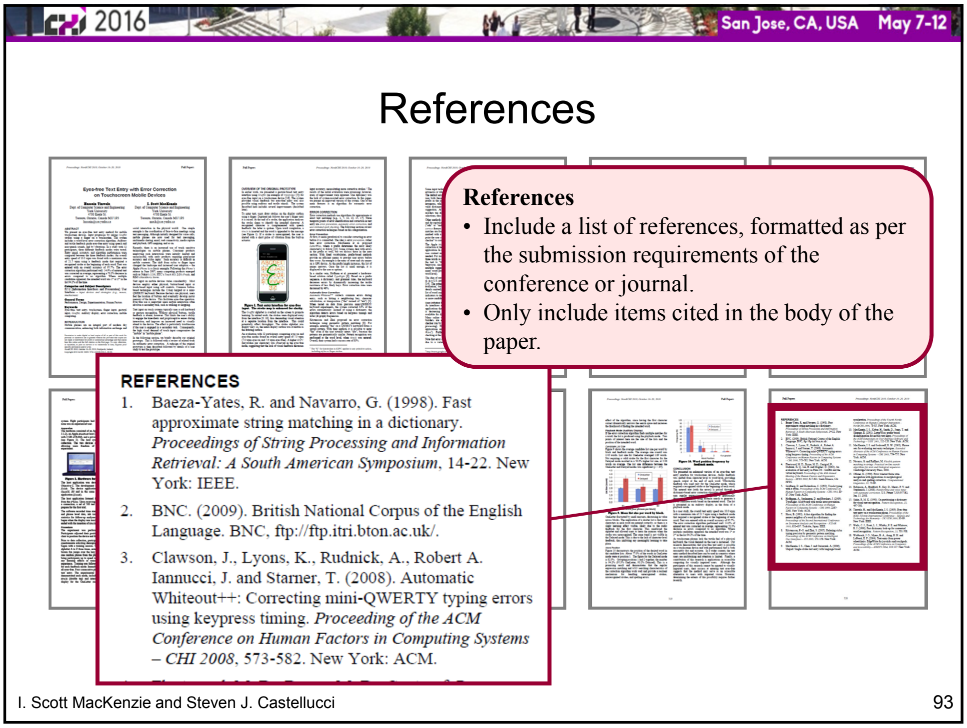
CHI 2016 San Jose, CA, USA May 7-12

Acknowledgment

Acknowledgment

- Optional.
- Thank people who helped with the research.
- Thank funding agencies.

I. Scott MacKenzie and Steven J. Castellucci 92



CHI 2016 San Jose, CA, USA May 7-12

Summary

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

Thank you

http://www.yorku.ca/mack/CHI2016-CourseNotes.zip

I. Scott MacKenzie and Steven J. Castellucci 94

Simple Experiment - Procedure

Attendees at a workshop, tutorial, or class are divided into groups of two.

Each group is given a copy of the two pages that follow.

The only materials needed are a timing device, such as a watch with a second hand, and a stylus, such as a ball-point pen with a cap or retractable tip.

The groups are asked to review the instructions at the top of the next page.

Initially, one partner serves as the "investigator", the other as "participant". Then the roles are reversed.

The first participant begins with "Method A".

The phrase is entered (see instructions) by the participant and timed by the investigator.

The time is recorded in the log sheet and the phrase is entered again.

Repeat for a total of five trials.

The participant switches to "Method B" and again enters the phrase five times, with entry for each phrase timed and logged.

Then, the partners switch roles. The participant becomes the investigator, and vice versa.

The order of keyboard entry is reversed for the second participant: "Method B" first, "Method A" second.

Demographic and experience data are also gathered on the log sheet.

The data are transcribed to the spreadsheet provided with this course (details to follow).

Results are immediately available. Discussion follows.

Instructions and Apparatus

Study and memorize the phrase below. Enter it by tapping with a non-marking stylus on the keyboard image. Proceed as quickly as possible while trying not to make mistakes. Don't forget to tap SPACE between words. Your partner will time you with a watch. Begin when your partner says "start". So that your partner knows when you finish, please say "stop" when you tap the last character (the "g" in "dog").

Enter the phrase five times using Method A, then five times using Method B. Then, switch roles with your partner. **Your partner should do Method B first, Method A second.**

Method "A"

<table>
<tr><td>Q</td><td>F</td><td>U</td><td>M</td><td>C</td><td>K</td><td>Z</td></tr>
<tr><td colspan="2">space</td><td>O</td><td>T</td><td>H</td><td colspan="2">space</td></tr>
<tr><td>B</td><td>S</td><td>R</td><td>E</td><td>A</td><td>W</td><td>X</td></tr>
<tr><td colspan="2">space</td><td>I</td><td>N</td><td>D</td><td colspan="2">space</td></tr>
<tr><td>J</td><td>P</td><td>V</td><td>G</td><td>L</td><td>Y</td><td></td></tr>
</table>

`the quick brown fox jumps over the lazy dog`

Method "B"

Q	W	E	R	T	Y	U	I	O	P
A	S	D	F	G	H	J	K	L	
Z	X	C	V	B	N	M			
space									

`the quick brown fox jumps over the lazy dog`

Log Sheet

Participant Initials:__________ Sex: Male Female Age:_________

Is English your first language? Yes No

Hours of computer use per day: __________

Do you regularly use a mobile phone? Yes No

Do you send text messages on a mobile phone? Yes No

If "yes", how many messages per day: __________

Method 'A' (first)	
Trial	Time
1	
2	
3	
4	
5	

Method 'B' (second)	
Trial	Time
1	
2	
3	
4	
5	

Participant Initials:__________ Sex: Male Female Age: _________

Is English your first language? Yes No

Hours of computer use per day: __________

Do you regularly use a mobile phone? Yes No

Do you send text messages on a mobile phone? Yes No

If "yes", how many messages per day: __________

Method 'A' (second)	
Trial	Time
1	
2	
3	
4	
5	

Method 'B' (first)	
Trial	Time
1	
2	
3	
4	
5	

Anova2

Anova2 - program to perform an analysis of variance on a table of data read from a file.

Five experimental designs are supported:

- One-way with one within-subjects factor
- One-way with one between-subjects factor
- Two-way with two within-subjects factors
- Two-way with one within-subjects factor and one between-subjects factor
- Three-way with two within-subjects factors and one between-subjects factor

Note on terminology: A "within-subjects factor" is often called a "repeated-measures factor". A "factor" is often called an "independent variable". The levels of a factor are called the "test conditions".

The data must be organized as a matrix with *p* rows and *n* columns. *p* is the number of participants (one per row) and *n* is the number of within-subjects test conditions (one per column). Each entry in the matrix contains a measurement on the behaviour of interest (e.g., entry speed or accuracy).

For designs with one within-subjects factor, *n* is the number of levels of the factor. For designs with two within-subjects factors, *n* is the product of the number of levels of each factor. For example, a two-factor experiment with repeated measures on 15 participants, having 2 levels on the first factor and 3 levels on the second, requires a data file with 15 rows and 2 x 3 = 6 columns. In total, there are 6 test conditions. Such an experiment is called a "2 x 3 within-subjects design".

If two within-subjects factors are used, the nesting of data is important. The columns are ordered with the levels of the second factor nested within the levels of the first factor. As an example, for a 2 x 3 design, the respective order of the data from columns one to six is F1L1-F2L1, F1L1-F2L2, F1L1-F2L3, F1L2-F2L1, F1L2-F2L2, F1L2-F2L3, where F = factor and L = level. The figure to the right illustrates.

Participant	F1L1	F1L1	F1L1	F1L2	F1L2	F1L2
	F2L1	F2L2	F2L3	F2L1	F2L2	F2L3
P01						
P02						
P03						
P04						
P05						
P06						
P07						
P08						
P09						
P10						
P11						
P12						

If a between-subjects factor is used, it appears as an additional column of nominal data. This is typically a group identifier, and is used, for example, if the participants were divided into groups to counterbalance the order of administering the within-subjects test conditions. The nominal data entries must consecutively for each group. Also, there must be the same number of participants in each group.

Usage:

```
java Anova2 file p f1 f2 f3 [-a] [-d] [-m] [-h]

   file = data file (comma or whitespace delimited)
   p = number of rows (participants) in data file
   f1 = number of levels for 1st within-subjects factor ("." if not used)
   f2 = number of levels for 2nd within-subjects factor ("." if not used)
   f3 = number of levels for between-subjects factor ("." if not used)
   -a = output ANOVA table
   -d = output debug data
   -m = output main effect means
   -h = data file includes header lines (see API for details)
   (Note: default is no output)
```

If a between-subjects factor is present, the `f3` argument is an integer corresponding to the number of groups. `f3`, if present, must divide evenly into `p`. For example, `p` = 15, `f1` = 3, and `f3` = 3 means the experiment involved 15 participants and a within-subjects factor with 3 levels. There will be 3 columns of data for the levels of the `f1` factor. Counterbalancing was used with participants divided into 3 groups, with 5 participants per group. (Here, "group" is treated like a between-subjects factors.) The groups are identified in the 4th column. The figure below to the right illustrates this. If a between-subjects factor is the only factor (e.g., gender), then the data file contains just two columns, one for the data and one to identify the groups. If any factor is not present, its argument is replaced with ".".

Participant	F1L1	F1L2	F1L3	Group
P01				1
P02				1
P03				1
P04				1
P05				1
P06				2
P07				2
P08				2
P09				2
P10				2
P11				3
P12				3
P13				3
P14				3
P15				3

-a option. The "`-a`" option produces the ANOVA table. The default is no output, so make sure either the `-a`, `-d`, or `-m` option is present.

-d option. The "`-d`" option produces a detailed output showing the original data as well as the means, sums of squares, mean squares, degrees of freedom, *F* statistics, and *p* for the *F* statistics, for all effects. Output is produced even for effects that do not exist, so don't fret if you see some output values coded as `NaN` or `Infinity`.

-m option. The "`-m`" option outputs the main effect means. This is useful to ensure that the data are properly extracted from the data matrix in computing the *F* statistics for the various main effects and interactions. For example, if the data are improperly nested for two-factor designs, this error will be apparent by comparing the output from this option against a manual calculation of the effect means.

-h option. The "`-h`" option is used if the data file contains header lines. In this case, the data file must have four header lines preceding the data, formatted as follows:

```
DV: <dependent_variable_name>
F1: <f1_name>, <f1_level_1_name>, <f1_level_2_name>, ...
F2: <f2_name>, <f2_level_1_name>, <f2_level_2_name>, ...
F3: <f3_name>
```

This option is strictly cosmetic. The output ANOVA table (`-a` option) and main effect means (`-m` option) will identify the dependent variable and the names and levels of the factors, as appropriate.

An example for each supported experiment design follows. For comparison, each analysis is also shown using a commercially available statistics package called *StatView* (currently available as *JMP*; http://www.jmp.com). The named data files are contained in the zip file containing the Anova2 application and API.

ONE-WAY WITH ONE WITHIN-SUBJECTS FACTOR

The file `dix-example-10x2.txt` contains

```
656,702
259,339
612,658
609,645
1049,1129
1135,1179
542,604
495,551
905,893
715,803
```

The data are hypothetical and appear in an example in Dix et al.'s *Human-Computer Interaction* (Prentice Hall, 2004, 3rd ed., p. 337). The single factor (F1) is Icon Design with two levels: Natural and Abstract. The data entries are the measurements on the dependent variable Task Completion Time (seconds). The data in the first column are the task completion time measurements for the Natural icons, while the data in the second column are the measurements for the Abstract icons. Each row contains the measurements taken on one participant. The hypothetical experiment used 10 participants.

The mean task completion times (not shown) are 697.7 s for the Natural icons and 750.3 s for the Abstract icons. An analysis of variance determines if there is a statistically significant difference between these means or if the difference is likely due to chance. The analysis is performed using

```
java Anova2 dix-example-10x2.txt 10 2 . . -a
```

and produces the following output on the console:

```
======================================================================
Effect          df            SS            MS          F        p
----------------------------------------------------------------------
Participant      9       1231492.000    136832.444
F1               1         13833.800     13833.800    33.359    3.0E-4
F1 x Par         9          3732.200       414.689
======================================================================
```

As seen in the table, and as might appear in a research paper, "The experiment revealed a significant effect of Icon Type on Task Completion Time ($F_{1,9}$ = 33.36, p < .0005)." Even though p = 3.0E-4 = .0003 in the ANOVA table, it is reported in research papers as $p < n$, where n is the closest more conservative value from the set .05, .01, .005, .001, .0005, .0001. Note also that in North American publications, the zero preceding the decimal point is typically omitted (because p is constrained between 0 and 1).

Two other outcomes are worth noting, where the results are non-significant. If p is above .05 and F > 1, p is reported as "p > .05". This means there is a greater than 5% chance that the difference in the means is due to chance. This is sufficient lack of confidence to deem the difference in the means "not significant". If p is above .05 and F <= 1, then p is not reported at all, but is replaced with "ns" meaning "not significant". This format is used is because it is impossible for differences in the means to be significant where F <= 1.

The results above are shown below in an ANOVA on the same data using *StatView*.

ANOVA Table for Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	1231492.000	136832.444				
Icon Type	1	13833.800	13833.800	33.359	.0003	33.359	1.000
Icon Type * Subject	9	3732.200	414.689				

Lambda and *Power* are not calculated in Anova2. *Lambda* is a measure of the noncentrality of the F distribution, calculated as F^2 x N, where N is the effect size (in this case, N is the degrees of freedom of the effect). *Power*, which ranges from 0 to 1, is the ability to detect an effect, if there is one. The closer to one, the more the experiment is likely to find an effect, if one exists in the population. *Power* > .80 is generally considered acceptable; i.e., if p is significant and *Power* > .80, then it is likely that the effect found actually exists.

TWO-WAY WITH ONE WITHIN-SUBJECTS FACTOR AND ONE BETWEEN SUBJECTS FACTOR

The hypothetical experiment described by Dix et al., because it was a within-subjects design, would like be design to use counterbalancing to cancel the learning effects that might occur as participants advanced from the first test condition to the second. Half the participants would be tested on the Natural icons first followed by the Abstract icons, while the other half would be tested in the reverse order. Like this, "Group" is a between-subjects factor with five participants in each group. To include this in the analysis, we append a column to the data file, creating a new data file called `dix-example-h10x2b.txt`. The new column identifies the groups as either "NA" (Natural first, Abstract second) or "AN" (Abstract first, Natural second). The file is also modified to include header lines, as per the requirements of the `-h` option. Here are the data:

```
DV: Completion Time (s)
F1: Icon Type, Natural, Abstract
F2: .
F3: Group
656,702,NA
259,339,NA
612,658,NA
609,645,NA
1049,1129,NA
1135,1179,AN
542,604,AN
495,551,AN
905,893,AN
715,803,AN
```

Note that the data for each group are in consecutive rows: 5 rows for the NA group, followed by 5 rows for the AN group. THIS IS IMPORTANT!

To analyse these data, we replace "." with "2" for `f3`, indicating the presence of a between-subjects factor with 2 groups. The `-h` option is also needed:

```
java Anova2 dix-example-h10x2b 10 2 . 2 -h -a

ANOVA table for Completion Time (s)
==========================================================================
Effect              df           SS            MS          F         p
--------------------------------------------------------------------------
Group               1        67744.800     67744.800     0.466     0.5142
Participant(Group)  8      1163747.200    145468.400
Icon Type           1        13833.800     13833.800    30.680     4.0E-4
Icon Type x Group   1          125.000       125.000     0.277     0.6128
Icon Type x P(Grou  8         3607.200       450.900
==========================================================================
```

The Group effect was not statistically significant ($F_{1,8}$ = 0.466, ns). This is good news, since it means counterbalancing worked; i.e., any learning effect that might have occurred for the AN group was effectively offset by a similar opposing learning effect for the NA group. The Icon Type (F1) x Group interaction effect also failed to achieve statistical significance ($F_{1,8}$ = 0.277, ns). This means there was no asymmetrical transfer of skill, also good news. Asymmetric transfer of skill means some aspect of one condition helped (or hindered!) the other condition, without a corresponding reverse effect. The same analysis in *StatView* appears as

ANOVA Table for Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Group	1	67744.800	67744.800	.466	.5142	.466	.091
Subject(Group)	8	1163747.200	145468.400				
Icon Type	1	13833.800	13833.800	30.680	.0005	30.680	.999
Icon Type * Group	1	125.000	125.000	.277	.6128	.277	.074
Icon Type * Subject(Group)	8	3607.200	450.900				

TWO-WAY WITH TWO WITHIN-SUBJECTS FACTORS

For the next example, we use data from an experiment on "eye typing" – the use of eye tracking technology for text entry using an on-screen soft keyboard. The experimental methodology and data analyses are described in a two-page short paper published in the *CHI 2003* proceedings, "Audio and visual feedback during eye typing". The paper is available in PDF format in the ACM Digital Library or in HTML format at

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

The experiment was a 4 x 4 repeated-measures design with 13 participants. The factors and levels were as follows:

Factor	Levels
Feedback Mode	Speech Only, Click+Visual, Speech+Visual, Visual Only
Block	B1, B2, B3, B4

There was no counterbalancing as the order of presenting the feedback modes was randomized. There were several dependent variables, however only error rate (%) is presented here. The error rate ANOVA table is in `errorrate-h13x16.txt`:

```
DV: Error Rate (%)
     F1: Feedback, Speech Only, Click+Visual, Speech+Visual, Visual Only
     F2: Block, B1, B2, B3, B4
     F3: .
     0.0,0.0,0.0,0.0,0.0,0.7142857142000001,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
     0.0,0.0,0.0,0.0,1.7391,0.0,0.0,0.0,0.0,0.9090,0.0,1.7391,1.9806,0.0,0.8695,0.0
     0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,4.4999,3.6363,0.0,0.0,0.9523,3.1111,0.0,0.0
     0.0,0.0,0.0,0.0,1.9047,0.0,0.0,1.8253,0.0,2.0202,0.0,0.0,8.5714,0.0,0.8695,0.0
     0.0,0.0,0.0,0.0,1.1111,1.1111,0.0,0.0,0.0,0.0,0.0,0.9090,1.4285,0.0,0.0,0.0
     0.0,0.0,0.0,0.0,1.0526,0.0,0.0,2.2222,0.0,0.0,1.7669,2.4319,0.0,0.7142,0.0,0.0
     0.9090,0.0,0.0,0.0,0.7407,1.7391,1.1111,0.0,0.8695,0.0,0.0,0.0,0.7407,1.8518,1.1111,0.0
     0.0,0.0,0.0,0.0,0.9090,0.0,0.0,0.0,0.0,0.0,0.0,0.7142,1.3793,0.0,0.7407,0.0
     0.0,0.0,0.0,0.8695,1.4814,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.1904,0.0,0.0,0.0
     0.0,0.0,0.0,0.0,0.7407,0.0,0.0,1.8461,1.0526,0.0,0.0,0.0,0.0,1.7669,1.4814,0.0
     0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7407,0.0,0.0,1.1111,0.7407,0.0,0.0
     1.6103,0.0,0.0,0.0,2.0,0.0,1.0526,0.7407,0.0,1.0526,0.0,0.0,2.0689,0.0,0.9090,0.8695
     4.3506,0.0,1.0526,0.0,1.3793,3.0526,0.0,1.8461,0.9523,0.0,2.714,0.0,8.063,0.0,5.497,1.481
```

The data entries in the file are full precision. The precision is reduced above to improve the presentation. Four header lines were manually inserted to improve the output generated by Anova2.

The `-m` option may be used prior to the analysis, to view the overall effect means:

```
java Anova2 errorrate-h13x16.txt 13 4 4 . -h -m
=============================
----- MAIN EFFECT MEANS -----
=============================
Grand mean: 0.5510756505634619
Participant means:
   p1=0.0446
   p2=0.4523
   p3=0.7625
   p4=0.9495
   p5=0.2850
   p6=0.5117
   p7=0.5671
   p8=0.2340
   p9=0.2213
   p10=0.4305
   p11=0.1620
   p12=0.6440
   p13=1.8994
Feedback means:
   Speech Only=0.1691
   Click+Visual=0.5831
   Speech+Visual=0.5002
   Visual Only=0.9520
Block means:
   B1=1.05367
   B2=0.44541
   B3=0.36877
   B4=0.33646
```

This is followed with an analysis of variance:

```
java Anova2 errorrate-h13x16.txt 13 4 4 . -h -a
```

```
ANOVA table for Error Rate (%)
==========================================================================
Effect                 df        SS            MS          F         p
--------------------------------------------------------------------------
Participant            12        43.903        3.659
Feedback                3        16.133        5.378       5.018     0.0052
Feedback x Par         36        38.577        1.072
Block                   3        17.839        5.946       5.070     0.0050
Block x Par            36        42.222        1.173
Feedback x Block        9        16.533        1.837       1.771     0.0820
Feedback x Block x Par 108      112.012        1.037
==========================================================================
```

As seen in the table, the main effect of Feedback Mode on Error Rate was significant ($F_{3,36}$ = 5.02, $p < .01$). There was also a significant improvement in entry speed with practice as evident by the significant effect of Block ($F_{3,36}$ = 5.07, $p < .001$). However, the Feedback Mode x Block interaction effect was not significant ($F_{9,108}$ = 1.771, $p > .05$).

The same data similarly analysed in *StatView* yield the following ANOVA table:

ANOVA Table for Error Rate (%)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	12	44.105	3.675				
Feedback	3	15.895	5.298	4.918	.0058	14.753	.885
Feedback * Subject	36	38.786	1.077				
Block	3	17.541	5.847	4.980	.0054	14.939	.890
Block * Subject	36	42.271	1.174				
Feedback * Block	9	16.144	1.794	1.727	.0914	15.542	.758
Feedback * Block * Subject	108	112.185	1.039				

THREE-WAY WITH TWO WITHIN-SUBJECTS FACTORS AND ONE BETWEEN-SUBJECTS FACTOR

The file `softkeyboard-h12x10b.txt` contains the data from an experiment to compare two layouts of soft keyboards. The experiment used 12 participants in a 2 x 5 repeated-measures design. The participants tapped the phrase "the quick brown fox jumps over the lazy dog" five times on each of two soft keyboard layouts. Each entry of a phrase is called a "trial". The dependent variable was Entry Speed in words per minute. There were two independent variables, or factors:

Factor	Levels
Layout	Opti, Qwerty
Trial	T1, T2, T3, T4, T5

Testing was counterbalanced: Each participant entered the phrase five times with one layout, then five times with the other layout. Half the participants used Opti first, following by Qwerty. The other half used the layouts in the reverse order. Thus, Group was a between-subjects factor with two levels, Group A and Group B.

The data file was edited to show the variable names in header lines and the participant group as the last entry on each data line:

```
DV: Entry Speed (wpm)
F1: Layout, Opti, Qwerty
F2: Phrase, P1, P2, P3, P4, P5
F3: Group
7.589,12.377,12.935,12.750,15.645,22.862,27.100,31.348,31.559,33.813,A
9.324,12.9,12.201,17.768,15.167,24.795,29.218,28.891,29.947,35.030,A
9.207,9.504,14.129,10.867,15.083,19.922,25.356,24.145,27.952,28.104,A
7.158,7.754,8.720,9.998,9.014,19.442,22.691,25.024,20.656,24.807,A
9.532,13.180,14.634,18.441,15.109,24.145,21.938,28.587,30.750,30.550,A
9.290,11.666,10.507,12.05,14.621,20.235,23.833,26.973,24.712,28.746,A
9.417,9.194,13.210,16.951,16.655,21.455,25.710,27.215,28.043,28.320,B
5.347,7.188,6.699,6.863,7.321,15.061,18.527,20.314,19.289,20.395,B
14.179,15.109,16.049,16.769,17.873,26.987,27.967,30.714,31.6758,34.014,B
8.970,10.396,11.400,13.741,13.013,20.756,24.783,27.682,25.519,25.121,B
9.552,12.693,16.602,17.725,20.722,22.0418,27.697,30.972,32.330,33.725,B
8.510,12.112,14.106,13.543,15.131,24.362,27.301,33.0769,32.845,32.0496,B
```

The data entries in the file are full precision. The precision is reduced above to improve the presentation.

The main effect means are computed as follows:

```
java Anova2 softkeyboard-h12x10b.txt 12 2 5 2 -h -m
============================
----- MAIN EFFECT MEANS -----
============================
Grand mean: 19.37750420382501
Participant means:
   p1=20.7983
   p2=21.5246
   p3=18.4274
   p4=15.5269
   p5=20.6872
   p6=18.2638
   p7=19.6175
   p8=12.7009
   p9=23.1342
   p10=18.1388
   p11=22.4065
   p12=21.3041
Layout means:
   Opti=12.2699
   Qwerty=26.4852
Trial means:
   T1=15.4229
   T2=18.1753
   T3=20.2560
   T4=20.9482
   T5=22.0851
Group means:
   A=19.2047
   B=19.5503
```

Entry speed in words per minute was much faster with the Qwerty layout (26.5 wpm) than with the Opti layout (12.3 wpm). Let's see if the variances were sufficiently low to deem the difference in the means statistically significant:

```
java Anova2 softkeyboard-h12x10b.txt 12 2 5 2 -h -a

ANOVA table for Entry Speed (wpm)
==========================================================================
Effect                  df        SS            MS           F         p
--------------------------------------------------------------------------
Group                    1        3.583         3.583      0.037    0.8523
Participant(Group)      10      981.236        98.124
Layout                   1     6062.238      6062.238    797.018    0.0000
Layout x Group           1        1.903         1.903      0.250    0.6278
Layout x P(Group)       10       76.062         7.606
Trial                    4      663.696       165.924     50.663    0.0000
Trial x Group            4        3.644         0.911      0.278    0.8904
Trial x P(Group)        40      131.003         3.275
Layout x Trial           4       22.466         5.617      2.707    0.0437
Layout x Trial x Group   4        6.856         1.714      0.826    0.5165
Layout x Trial x P(Gro  40       83.003         2.075
==========================================================================
```

Yes. The *F* statistic, which is the ratio of the mean squares (6062.238 / 7.606 = 797.0) is extremely high. Not surprisingly, the *F* statistic for the main effect of Layout on Entry Speed is highly significant ($F_{1,10} = 797.0$, $p < .0001$). In all, the table shows three main effects and four interaction effects. There is considerable leeway in presenting the results in a research paper. See Experiment 1 in Using paper mockups for evaluating soft keyboard layouts for an example of how the results above might be reported.

The results above are confirmed using *StatView*:

ANOVA Table for Entry Speed (wpm)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Group	1	3.583	3.583	.037	.8523	.037	.053
Subject(Group)	10	981.236	98.124				
Layout	1	6062.238	6062.238	797.018	<.0001	797.018	1.000
Layout * Group	1	1.903	1.903	.250	.6278	.250	.073
Layout * Subject(Group)	10	76.062	7.606				
Trial	4	663.696	165.924	50.663	<.0001	202.651	1.000
Trial * Group	4	3.644	.911	.278	.8904	1.113	.104
Trial * Subject(Group)	40	131.003	3.275				
Layout * Trial	4	22.466	5.617	2.707	.0437	10.827	.696
Layout * Trial * Group	4	6.856	1.714	.826	.5165	3.304	.235
Layout * Trial * Subject(Group)	40	83.003	2.075				

ONE-WAY WITH ONE BETWEEN-SUBJECTS FACTOR

A one-way between-subjects design might be used, for example, to test whether an interface or interaction technique works better with left-handed vs. right-handed users (or with males vs. females). In this case, the design must be between-subjects because a participant cannot be both left-handed and right-handed (or male and female!). Two groups of participants are required. Let's consider the case where five left-handed users (L) and five right-handed users (R) are measured on a task. The independent variable is Handedness with two levels, Left and Right, and the dependent variable is Time (seconds) to complete a task. Here are the example data, stored in `anova-h10b.txt`:

```
DV: Time (s)
F1: .
F2: .
F3: Handedness
25.6,L
23.4,L
19.4,L
28.1,L
25.9,L
14.3,R
22.0,R
30.4,R
21.1,R
19.3,R
```

The means (not shown) for the Left- and Right-handed groups were 28.48 s and 21.42 s, respectively. So, the Left-handed group took, on average, 33% longer to complete the task. That's a huge performance difference, but is the difference in the means statistically significant? Let's see. The analysis is performed as follows:

```
java Anova2 anova-h10b.txt 10 . . 2 -h -a

ANOVA table for Time (s)
==========================================================================
Effect              df            SS             MS          F        p
--------------------------------------------------------------------------
Handedness           1            23.409         23.409      1.043    0.3371
Residual             8           179.616         22.452
==========================================================================
```

Despite the observation that the Left-handed group took considerable longer to complete the task, the difference between the groups was not statistically significant ($F_{1,8} = 1.04$, $p > .05$). This might be partly attributed to the small number of participants tested. It might also be attributed simply to a lack of bias in the interface for Left-handed vs. Right-handed users.

Using *StatView*, the above results are confirmed:

ANOVA Table for Time(s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Handedness	1	23.409	23.409	1.043	.3371	1.043	.142
Residual	8	179.616	22.452				

A note on the calculations:

The trickiest part is the calculation of *p*, representing the significance of *F*. This comes by way of the method `FProbability` in the `Statistics` class in the University of Waikato's *Weka* package (http://www.cs.waikato.ac.nz/ml/weka/index.html).The source file was obtained by downloading the archive

http://sourceforge.net/projects/weka/files/weka-3-4/3.4/weka-3-4.jar

then extracting `\weka-3-4\weka-src.jar\weka\core\Statistics.java` using an unzip program.

Author:

Scott MacKenzie, 2003-2014

PostHoc

PostHoc - a Java utility for performing post hoc pairwise comparisons on a data set.

In experimental research, an analysis of variance (ANOVA), or *F*-test, investigates the effect of an independent variable on a dependent variable. An independent variable is a condition manipulated in an experiment, such as *input device* (*mouse* vs. *touchpad*). A dependent variable is a human performance measurement, such as task completion time.

An independent variable is manipulated over levels. In HCI, the levels of an independent variable are often called *test conditions*. In statistics references, they are often called *treatments*. An independent variable must have at least two levels, as in the mouse vs. touchpad example above. However, an independent variable often has more than two levels. An example might be *feedback mode* with levels *auditory*, *visual*, *tactile*, *none*. If the independent variable has more than two levels, a significant *F*-test only indicates that at least one level is significantly different from one other level. It does not indicate which levels differ significantly from one another. To determine which levels (test conditions) different significantly from one another, a post hoc pairwise comparisons test is used.

This utility supports four post hoc comparisons tests:

- Fisher's LSD test
- Bonferroni-Dunn test
- Tukey's HSD test
- Scheffé test

Invocation (usage message if invoked without arguments):

```
Usage: java PostHoc file [-bn] [-g] [-hn] [-v] [-flsd] [-bd] [-thsd] [-sch] [-all]

  where 'file' contains a data table (comma or space delimited)
        [-bn] = data organized vertically in blocks of 'n' scores
                (NOTE: 'n' must divided evenly into the number of rows of data)
        [-g]  = group code in right-hand column (will be ignored)
        [-hn] = ignore n header lines (e.g., use '-h4' to ignore 1st 4 lines)
        [-v]  = verbose (debugging and other information)
        [-flsd] = output Fisher LSD post hoc results
        [-bd]   = output Bonferroni-Dunn post hoc results
        [-thsd] = output Tukey HSD post hoc results
        [-sch]  = output Scheffe post hoc results
        [-all]  = output all post hoc results (see above)
        NOTE: default is no output
```

The command-line options accommodate data in a variety of formats. If the data are organized in a rectangular table, then command-line options are not needed except for the name of the data file and one of the output options. The assumption in this case is that the data are organized in an $n \times k$ table, with n rows and k columns. n is the number of participants, k is the number of test conditions. This is the usual format for data in a within-subjects experiment. With k conditions there are $k \times (k - 1) / 2$ comparisons.

If the file contains header lines, use the `-hn` option. The first n lines in the data file are ignored.

For between-subjects experiments, the data are sometimes organized vertically in a single column. In this case, use the `-bn` option with n identifying the number of participants per group. Note that n must divide evenly into the total number of rows of data. So, with $n = 6$ and 18 rows of data (each of length 1), the experiment has $18 / 6 = 3$ conditions with 6 participants tested on each condition. The first 6 rows are the scores for the first condition, the next 6 for the second condition, and the last 6 for the third condition.

Between-subjects data sets often include a categorical group identifier in the last column. In this case, use the `-g` option. The group identifier is ignored.

The primary source for coding these tests was Chapter 21 in D. J. Sheskin's *Handbook of Parametric and Nonparametric Statistical Procedures*, 5th ed., CRC Press, 2011, pp. 886-999. The chapter begins with an overview of the between-subjects analysis of variance. The following data set is used as an example throughout the chapter:

```
 8      7      4
10      8      8
 9      5      7
10      8      5
 9      5      7
```

There are 3 groups of participants with 5 participants in each group. Each group was tested on a different condition, with the scores for each condition/group appearing in separate columns. There were 15 total participants.

As well as describing the calculations leading to the *F*-test, Sheskin demonstrates the various post hoc comparisons tests that often follow a statistically significant result in the ANOVA. The calculations are embedded in this utility for the four tests noted above. The following illustrates:

```
PROMPT>type sheskin-ex1.txt
 8     G1
10     G1
 9     G1
10     G1
 9     G1
 7     G2
 8     G2
 5     G2
 8     G2
 5     G2
 4     G3
 8     G3
 7     G3
 5     G3
 7     G3

PROMPT>java PostHoc sheskin-ex1.txt -b5 -g -v -all
n = 5 (number of participants or participants/group)
N = 15 (n * k)
k = 3 (number of groups/conditions)
Raw data...
    8.00    7.00    4.00
   10.00    8.00    8.00
    9.00    5.00    7.00
   10.00    8.00    5.00
    9.00    5.00    7.00
groupTotals = 46.0, 33.0, 31.0,
groupMeans = 9.2, 6.6, 6.2,
ssGroup = 426.0, 227.0, 203.0,
dfBG = 2
dfWG = 12
dfTotal = 14
ssTotal = 49.33333333333337
ssBG = 26.533333333333417
ssWG = 22.80000000000001
msBG = 13.266666666666708
msWG = 1.900000000000001
F = 6.982456140350895
p = 0.009744651634258016

cdLSD = 1.8994399174493526 (critical difference for Fisher's LSD test)
----------------------------------------------------------
---------- Pairwise Comparisons (Fisher LSD) -------------
----------------------------------------------------------
Pair 1:2 -->     2.60   >     1.90   ?     * (significant)
Pair 1:3 -->     3.00   >     1.90   ?     * (significant)
Pair 2:3 -->     0.40   >     1.90   ?     -
----------------------------------------------------------
z = 2.3932468761578614
tBD = 2.795768899705345
cdBD = 2.4372948206619007 (critical difference for Bonferroni-Dunn test)
----------------------------------------------------------
-------- Pairwise Comparisons (Bonferroni-Dunn) ----------
----------------------------------------------------------
Pair 1:2 -->     2.60   >     2.44   ?     * (significant)
Pair 1:3 -->     3.00   >     2.44   ?     * (significant)
Pair 2:3 -->     0.40   >     2.44   ?     -
----------------------------------------------------------
```

```
cdHSD = 2.3258334033201957 (critical difference for Tukey-HSD test)
----------------------------------------------------------
----------- Pairwise Comparisons (Tukey-HSD) -------------
----------------------------------------------------------
Pair 1:2 -->     2.60   >     2.33   ?     * (significant)
Pair 1:3 -->     3.00   >     2.33   ?     * (significant)
Pair 2:3 -->     0.40   >     2.33   ?     -
----------------------------------------------------------
cdSCH = 2.430155550576959 (critical difference for Scheffe test)
----------------------------------------------------------
------------ Pairwise Comparisons (Scheffe) --------------
----------------------------------------------------------
Pair 1:2 -->     2.60   >     2.43   ?     * (significant)
Pair 1:3 -->     3.00   >     2.43   ?     * (significant)
Pair 2:3 -->     0.40   >     2.43   ?     -
----------------------------------------------------------
```

The verbose option is used, so the output is replete with additional information. The goal of identifying which conditions differ significantly from one another appears in the tables for the four tests. The results are consistent: Pairs 1:2 and 1:3 differ significantly.

The most difficult calculation is that for the critical difference. The values above are consistent with those in Sheskin. As further verification, the same test was undertaken with *StatView* (now JMP). The results appear in the figure to the right.

Fisher's PLSD for Score
Effect: Group
Significance Level: 5 %

	Mean Diff.	Crit. Diff	P-Value	
G1, G2	2.600	1.899	.0114	S
G1, G3	3.000	1.899	.0049	S
G2, G3	.400	1.899	.6546	

Bonferroni/Dunn for Score
Effect: Group
Significance Level: 5 %

	Mean Diff.	Crit. Diff	P-Value	
G1, G2	2.600	2.423	.0114	S
G1, G3	3.000	2.423	.0049	S
G2, G3	.400	2.423	.6546	

Comparisons in this table are not significant unless the corresponding p-value is less than .0167.

Tukey/Kramer for Score
Effect: Group
Significance Level: 5 %

	Mean Diff.	Crit. Diff	
G1, G2	2.600	2.324	S
G1, G3	3.000	2.324	S
G2, G3	.400	2.324	

Scheffe for Score
Effect: Group
Significance Level: 5 %

	Mean Diff.	Crit. Diff	P-Value	
G1, G2	2.600	2.430	.0359	S
G1, G3	3.000	2.430	.0163	S
G2, G3	.400	2.430	.9009	

Again, the results are consistent. There are minor deviations between some of the critical values, however. For example, the Bonferroni-Dunn critical value calculated by this utility is 2.44. The critical value in Sheskin's handbook is 2.43 (p. 907) and by *StatView* 2.423 (see above). The differences may be due to rounding or to the use of embedded lookup tables in *StatView* versus the use of methods of the `Statistics` class in this Java utility.

Significance is at the $\alpha = .05$ level. For post hoc comparisons, the alpha level is "family-wise". This implies that the overall alpha for the set of comparisons is $\alpha = .05$. To maintain the overall alpha, the individual comparisons use a more stringent alpha level. The Bonferroni-Dunn test, for example, uses $.05 / n$, where n is the number of comparisons. With three comparisons, $\alpha = .05 / 3 = .0167$. Thus, a comparison is only deemed significant if $p < .0167$. See above. The other tests have different approaches to maintaining the family-wise alpha level. See Sheskin for complete details.

Note that if the above data are re-organized into the within-subjects format for *StatView* (compact variables), the result of the pairwise comparisons test is the same. This is not as precise since the residual mean squares are different for within-subjects data than for between-subjects data. However, the differences are minor. Sheskin provides additional details.

MacKenzie's *Human Computer Interaction: An Empirical Research Perspective* includes an example in Chapter 6 of a post hoc comparisons test for data in a hypothetical experiment. The experiment was within-subjects with 16 participants tested on four conditions (A, B, C, and D). The ANOVA was statistically significant. With 4 test conditions, there are $4 \times (4 - 1) / 2 = 6$ pairwise comparisons possible. Which conditions differ from one another? This is tested as follows (without the `-v` option):

```
PROMPT>type posthoc-ex1.txt
11     11     21     16
18     11     22     15
17     10     18     13
19     15     21     20
13     17     23     10
10     15     15     20
14     14     15     13
13     14     19     18
19     18     16     12
10     17     21     18
10     19     22     13
16     14     18     20
10     20     17     19
10     13     21     18
20     17     14     18
18     17     17     14

PROMPT>java PostHoc posthoc-ex1.txt -all
----------------------------------------------------------
---------- Pairwise Comparisons (Fisher LSD) -------------
----------------------------------------------------------
Pair 1:2 -->     0.88   >     2.30   ?     -
Pair 1:3 -->     4.50   >     2.30   ?     * (significant)
Pair 1:4 -->     1.81   >     2.30   ?     -
Pair 2:3 -->     3.63   >     2.30   ?     * (significant)
Pair 2:4 -->     0.94   >     2.30   ?     -
Pair 3:4 -->     2.69   >     2.30   ?     * (significant)
----------------------------------------------------------
----------------------------------------------------------
-------- Pairwise Comparisons (Bonferroni-Dunn) ----------
----------------------------------------------------------
Pair 1:2 -->     0.88   >     2.83   ?     -
Pair 1:3 -->     4.50   >     2.83   ?     * (significant)
Pair 1:4 -->     1.81   >     2.83   ?     -
Pair 2:3 -->     3.63   >     2.83   ?     * (significant)
Pair 2:4 -->     0.94   >     2.83   ?     -
Pair 3:4 -->     2.69   >     2.83   ?     -
----------------------------------------------------------
----------------------------------------------------------
----------- Pairwise Comparisons (Tukey-HSD) -------------
----------------------------------------------------------
Pair 1:2 -->     0.88   >     3.03   ?     -
Pair 1:3 -->     4.50   >     3.03   ?     * (significant)
Pair 1:4 -->     1.81   >     3.03   ?     -
Pair 2:3 -->     3.63   >     3.03   ?     * (significant)
Pair 2:4 -->     0.94   >     3.03   ?     -
Pair 3:4 -->     2.69   >     3.03   ?     -
----------------------------------------------------------
----------------------------------------------------------
------------ Pairwise Comparisons (Scheffe) --------------
----------------------------------------------------------
Pair 1:2 -->     0.88   >     3.30   ?     -
Pair 1:3 -->     4.50   >     3.30   ?     * (significant)
Pair 1:4 -->     1.81   >     3.30   ?     -
Pair 2:3 -->     3.63   >     3.30   ?     * (significant)
Pair 2:4 -->     0.94   >     3.30   ?     -
Pair 3:4 -->     2.69   >     3.30   ?     -
----------------------------------------------------------
```

The results are mostly consistent. Pairs 1:3 (A:C) and 2:3 (B:C) differ significantly. Fisher's LSD test also deemed the pair 3:4 (C:D) significantly different. The results are the same with *StatView*. Explaining the slightly different outcomes for the various post hoc comparisons tests is to venture into territory that is outside the scope of this API. The interested reader is directed to the introductory comments for each test in Sheskin's handbook: Fisher's LSD test (p. 903), the Bonferroni-Dunn test (p. 906), Tukey's HSD test (p. 909), and the Scheffé test (p. 913).

Author:

Scott MacKenzie, 2012-2014

Suggested Readings on Empirical Research Methods

All material in this course is elaborated in greater detail in the following reference:

MacKenzie, I. S. (2013). *Human-computer interaction: An empirical research perspective*. Waltham, MA: Morgan Kaufmann.

Downloads supporting material in the book (and in this course) are found at

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

Additional references on topics pertinent to this course with citations to relevant papers or books are as follows:

- Empirical research methods [2, 12, 18]
- Use of paper mock-ups for research in HCI [1, 4-7, 10, 19-22, 24]
- Example papers by MacKenzie presenting user studies conforming to conventional HCI practices in empirical research [11, 13-17, 23]
- Longitudinal studies with learning curves showing participants' performance improvement with practice [3, 8, 14, 16, 25, 26]
- Experiments with a between-subjects factor and discussion on the rationale for such [9, 14]

Bibliography

1. Aliakseyeu, D. and Martens, J.-B., The electronic paper prototype with visual interaction enriched windows, *Proceedings of the 2nd European Union symposium on Ambient Intelligence*, (New York: ACM, 2004), 11-14.
2. APA, *Publication manual of the American Psychological Association*, 4th ed. Washington, DC: APA, 2009.
3. Bellman, T. and MacKenzie, I. S., A probabilistic character layout strategy for mobile text entry, *Proceedings of Graphics Interface '98*, (Toronto: Canadian Information Processing Society, 1998), 168-176.
4. Chandler, C. D., Lo, G., and Sinha, A. K., Multimodal theater: Extending low fidelity paper prototyping to multimodal applications, *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems - CHI '02*, (New York: ACM, 2002), 874-875.
5. Grady, H. M., Web site design: A case study in usability testing using paper prototypes, *Proceedings of the 18th Annual ACM Conference on Computer Documentation*, (New York: ACM, 2000), 39-45.
6. Hanington, B. M., Interface in form: Paper and product prototyping for feedback and fun, in *Interactions*, vol. 13: New York: ACM, 2006, January, 28-30.
7. Hendry, D. G., Mackenzie, S., Kurth, A., Spielberg, F., and Larkin, J., Evaluating paper prototypes on the street, *Extended Abstracts of the ACM Conference in Human Factors in Computing Systems - CHI '05*, (New York: ACM, 2005), 1447-1450.
8. Isokoski, P. and Raisamo, R., Quikwriting as a multi-device text entry method, *Proceedings of the Third Nordic Conference on Human-Computer Interaction - NordiCHI 2004*, (New York: ACM, 2004), 105-108.
9. Kabbash, P., MacKenzie, I. S., and Buxton, W., Human performance using computer input devices in the preferred and non-preferred hands, *Proceedings of the INTERCHI '93 Conference on Human Factors in Computing Systems*, (New York: ACM, 1993), 474-481.
10. Liu, L. and Khooshabeh, P., Paper or interactive?: A study of prototyping techniques for ubiquitous computing environments, *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems - CHI '03*, (New York: ACM, 2003), 1030-1031.
11. MacKenzie, I. S., Mobile text entry using three keys, *Proceedings of the Second Nordic Conference on Human-Computer Interaction - NordiCHI 2002*, (New York: ACM, 2002), 27-34.
12. MacKenzie, I. S., Evaluation of text entry techniques, in *Text entry systems: Mobility, accessibility, universality*, (I. S. MacKenzie, and Tanaka-Ishii, K., Ed.). San Francisco, Morgan Kaufmann, 2007, 75-101.

13. MacKenzie, I. S., Chen, J., and Oniszczak, A., Unipad: Single-stroke text entry with language-based acceleration, *Proceedings of the Fourth Nordic Conference on Human-Computer Interaction - NordiCHI 2006*, (New York: ACM, 2006), 78-85.
14. MacKenzie, I. S., Kober, H., Smith, D., Jones, T., and Skepner, E., LetterWise: Prefix-based disambiguation for mobile text entry, *Proceedings of the ACM Symposium on User Interface Software and Technology - UIST 2001*, (New York: ACM, 2001), 111-120.
15. MacKenzie, I. S. and Oniszczak, A., A comparison of three selection techniques for touchpads, *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems - CHI '98*, (New York: ACM, 1998), 336-343.
16. MacKenzie, I. S. and Zhang, S. X., The design and evaluation of a high-performance soft keyboard, *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems - CHI '99*, (New York: ACM, 1999), 25-31.
17. MacKenzie, I. S. and Zhang, S. X., An empirical investigation of the novice experience with soft keyboards, *Behaviour & Information Technology, 20*, 2001, 411-418.
18. Martin, D. W., *Doing psychology experiments*, 6th ed. Belmont, CA: Wadsworth, 2004.
19. Nielson, J., Paper versus computer implementations as mockup scenarios for heuristic evaluation, *Proceedings of IFIP INTERACT '90: Human-Computer Interaction*, (Berlin: Springer, 1990), 315-320.
20. Sefelin, R., Tscheligi, M., and Giller, V., Paper prototyping - what is it good for?: A comparison of paper- and computer-based low-fidelity prototyping, *Extended Abstract of the ACM Conference on Human Factors in Computing Systems - CHI '03*, (New York: ACM, 2003), 778-779.
21. Slaughter, I., Oard, D. W., Warnick, V. I., Harding, J. L., and Wilderson, G. J., A graphical interface for speed-based retrieval, *Proceedings of the 3rd ACM International Conference on Digital Libraries -- DL '98*, (New York: ACM, 1998), 305-306.
22. Snyder, C., *Paper prototyping: The fast and easy way to design and refine user interfaces*. San Francisco: Morgan Kaufmann, 2003.
23. 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.
24. Tohidi, M., Buxton, W., Baecker, R., and Sellen, A., Getting the right design and the design right: Testing many is better than one, *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems - CHI 2006*, (New York: ACM, 2006), 1243-1252.
25. Wigdor, D. and Balakrishnan, R., TiltText: Using tilt for text input to mobile phones, *Proceedings of the ACM Symposium on User Interface Software and Technology - UIST 2003*, (New York: ACM, 2003), 81-90.
26. Wobbrock, J. O. and Myers, B. A., Few-key text entry revisited: Mnemonic gestures on four keys, *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems - CHI 2006*, (New York: ACM, 2006), 489-492.