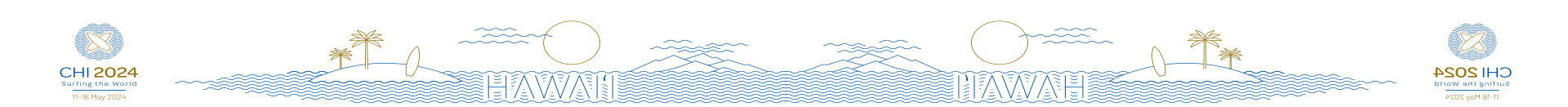


Empirical Research Methods for Human-Computer Interaction[1]

I. Scott MacKenzie[1]

with Janet Read[2] & Matt Horton[2]

[1]York University
Toronto, Canada

[2] U. of Central Lancashire
Preston, UK

CHI2024, May 11-16, Honolulu, USA.
ACM 2024/05.

Presenter

Scott 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. Scott is a member of the SIGCHI Academy. He has more than 200 HCI publications (including 50 from the SIGCHI conference and two HCI books) and has given numerous invited talks over the past 25 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/

Janet Read and **Matt Horton** have previously delivered courses at CHI on Child Computer Interaction. For the last 15 years Janet has been teaching a course on research methods and Matt has been teaching an advanced level course in user studies in HCI.

Full details: https://chici.org/about/

Topics

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

What is…Research

(three dictionary definitions)

1. Careful or diligent search

2. Collecting information about a subject

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

Definition #4
(not in dictionary)

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

Example (Definition #4)

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

Rogers Communications, Inc.

What is… *Empirical* Research

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

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
- In HCI…
  - The methodology is often a ***user study*** (an experiment with human participants)

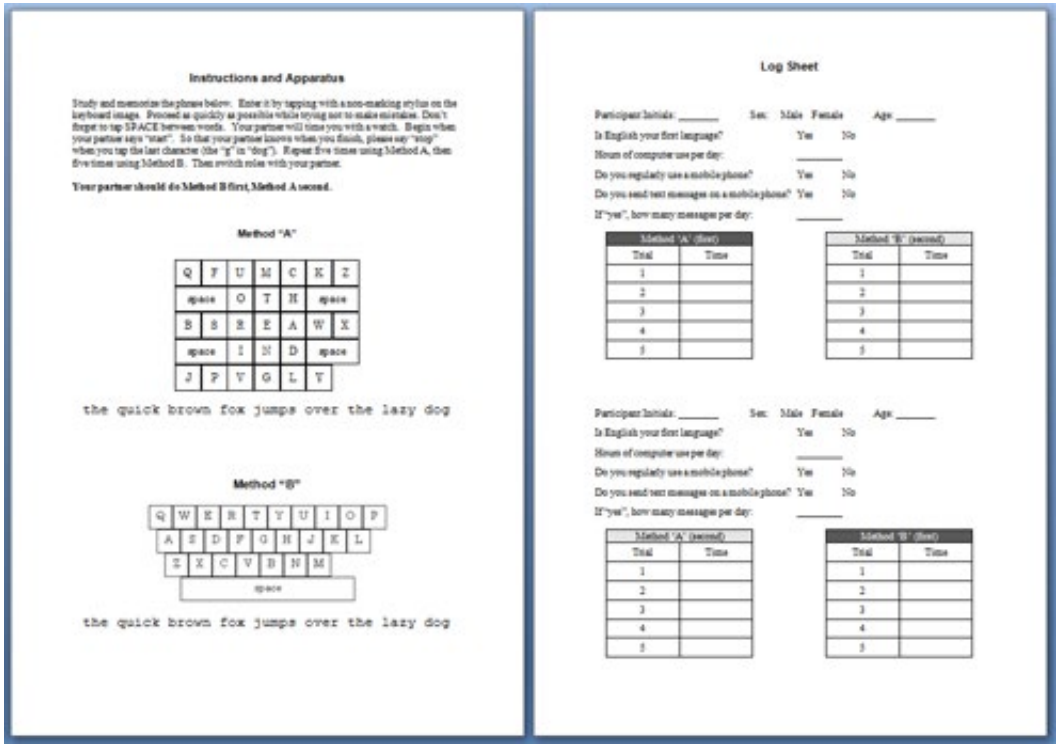
Topics

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

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

Handout (2 pages)



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"). Repeat 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"

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

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	

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

- Remember:

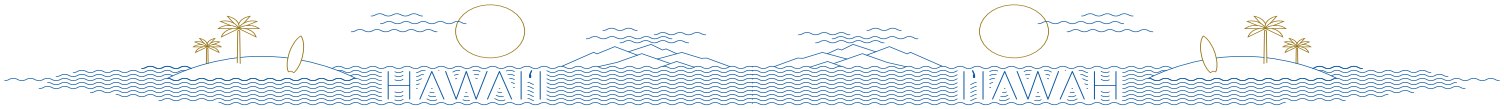
The second person to do the task needs to do *Method B first*, followed by Method A

Do the Experiment

- The experiment is performed
- This takes about 25 minutes
- Assistants transcribe the tabulated data into a ready-made spreadsheet
- Results are presented in Session Two

Topics

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



Observations and Measurements

- Observations are gathered…
  - Manually (human observers)
  - Automatically (computers, software, cameras, sensors, etc.)

- A measurement is a recorded observation

Scales of Measurement[1]

- Nominal
- Ordinal
- Interval
- Ratio

crude

sophisticated

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

Scales of Measurement

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

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

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

Real Data !

Note: The counts (grey) are ratio scale measurements

Scales of Measurement

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

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

What is your weekly time playing computer games?

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

Scales of Measurement

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

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

Indicate your level of agreement with the following statement:

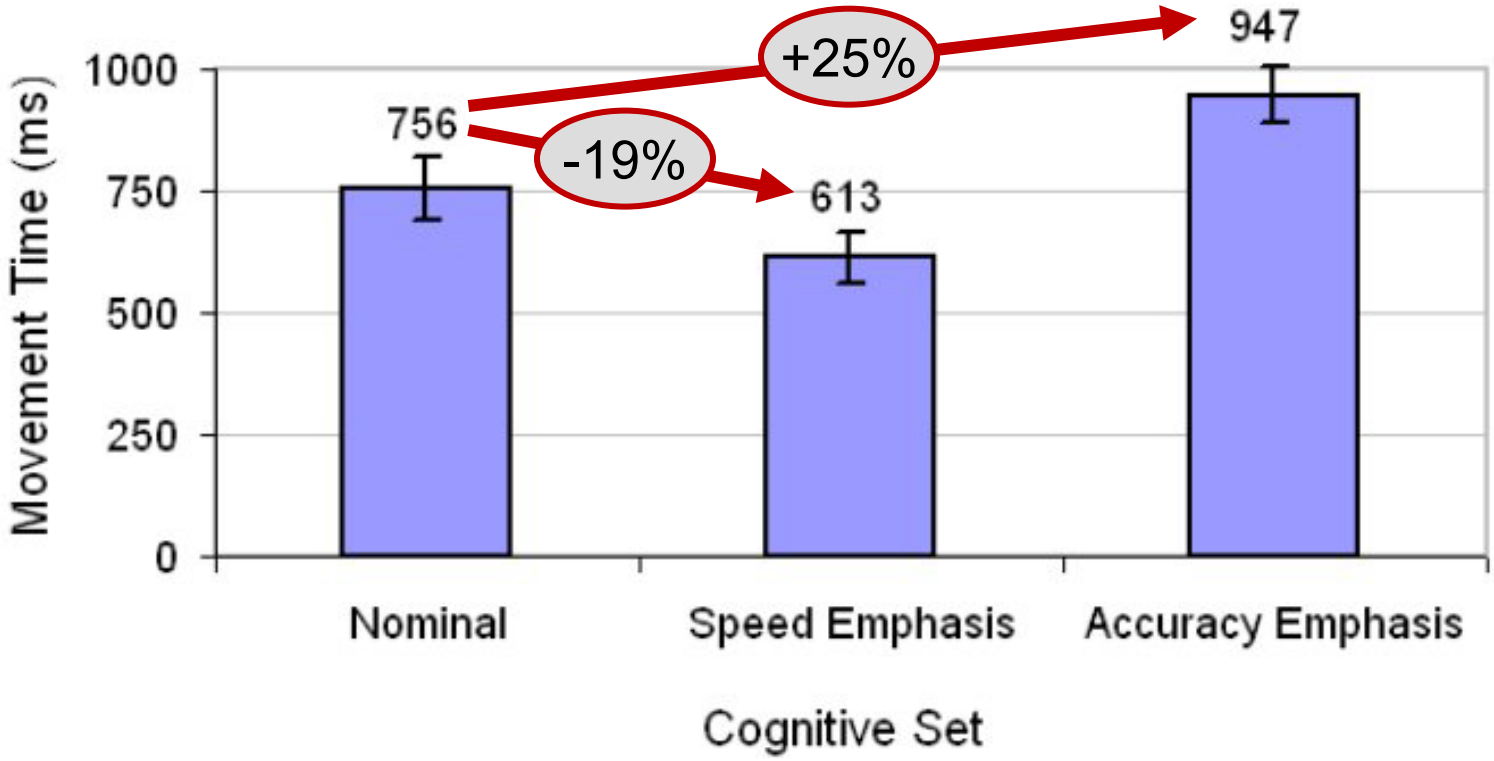
	Strongly disagree				Strongly agree
It is safe to talk on a mobile phone while driving.	1	2	3	4	5

Scales of Measurement

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

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

Ratio Data Example in HCI[1]



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

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

Topics

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

Allen Newell (1927-1992)

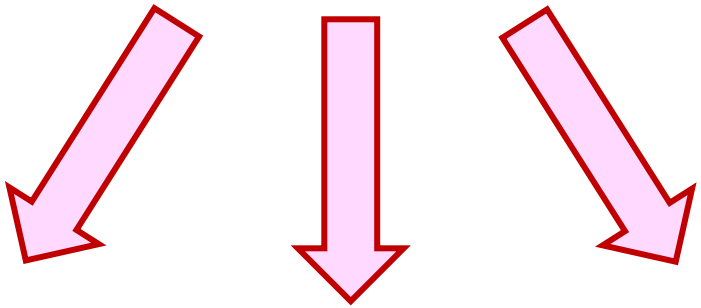
- ACM Turing Award (1975)
- With Stuart Card and Tom Moran, author of *The Psychology of Human-Computer Interaction* (1983)

"Science is method. Everything else is commentary."

Research Methods

Theoretical Research

Empirical Research



Observational Method

Correlational Method

Experimental Method

Observational Method

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

Experimental Method

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

Correlational Method

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

Research Methods

Observational

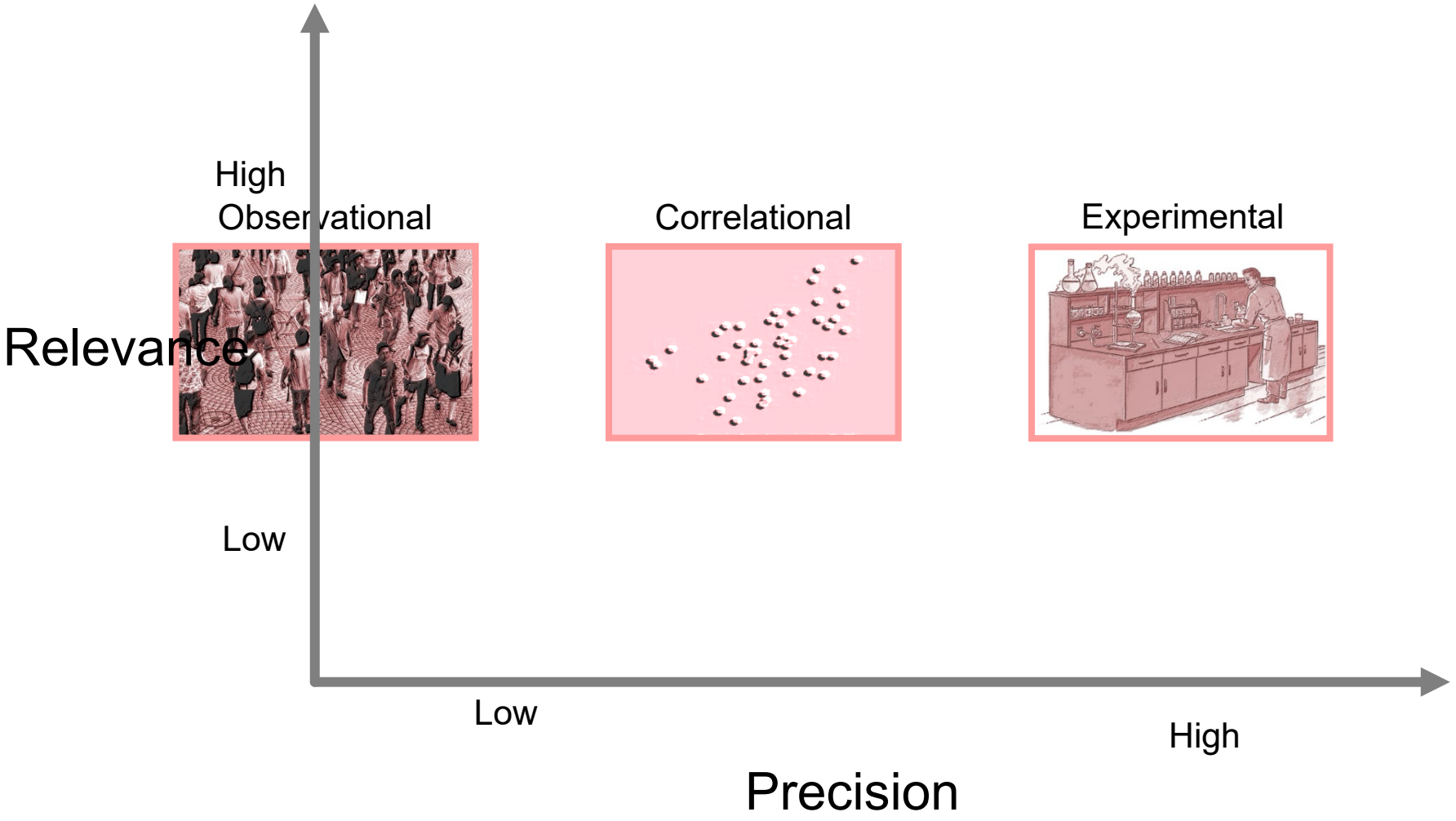
- Real-world setting
- No variables per se
- Broad, qualitative questions:
  - What's going on?
- High-level inquiry

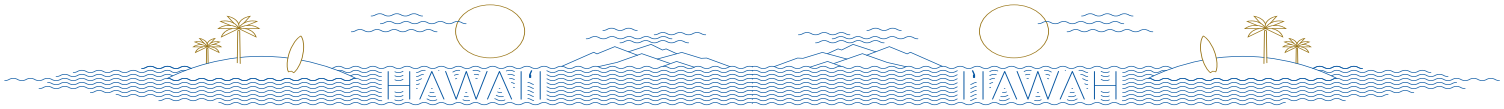
Correlational

Experimental

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

Relevance vs. Precision





Topics

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

Experiment Terminology (Part 1)

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

Participant

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

[1] APA. (2020). *Publication Manual of the American Psychological Association* (7th ed.) Washington, DC: APA, p. 73.

How Many Participants

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

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

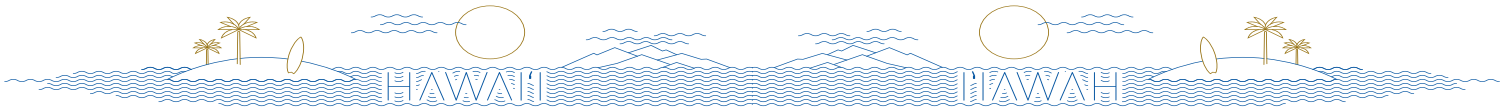
How Many Participants (Part 2)

- UI researchers in industry use about five participants when testing a new system… Why?
- Useful for exploratory testing of a UI
  - Five participants (UI experts) will discover most of the problems[1]
  - Diminishing returns if >5 participants used[1]
- Commonly called "usability evaluation"
- Not practical for a user study comparing two or more interaction techniques

[1] Nielsen et al. (1993). "A mathematical model of the finding of usability problems," *Proc. ACM INTERCHI'93*, pp. 206-213

Independent Variable

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



Test Conditions

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

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

Dependent Variable

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

Control Variable

- ***Control variable*** – a circumstance (not under investigation) that is held constant
- Upside: helps internal validity (better chance of obtaining statistical significance)
- Downside: hinders external validity (results are less generalizable to other people and other situations)
- Typical examples
  - Lighting
  - Room
  - Room temperature
  - Participant position (e.g., sitting)
  - Device location (e.g., on a desk)

Random Variable

- ***Random variable*** – a circumstance that is allowed to vary randomly
- Upside: aides external validity (results are more generalizable)
- Downside: hinders internal validity (more variability is introduced in the measures)
- Typical examples
  - Time since last meal
  - Coffee consumption prior to testing
  - Time of day for testing (e.g., morning, afternoon, evening)
  - Participants' field of study or work
  - Participants' socio-economic background

Tradeoff

(control variable vs. random variable)

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

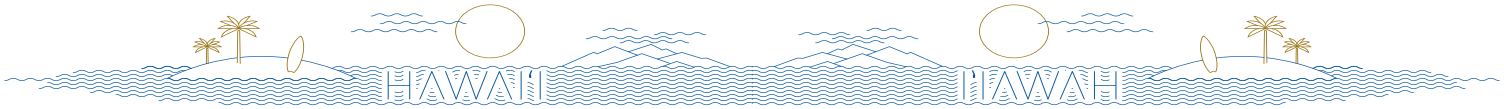
Variable	Advantage	Disadvantage
Random	Improves external validity by using a variety of situations and people.	Compromises internal validity by introducing additional variability in the measured behaviours.
Control	Improves internal validity since variability due to a controlled circumstance is eliminated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

Confounding Variable

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

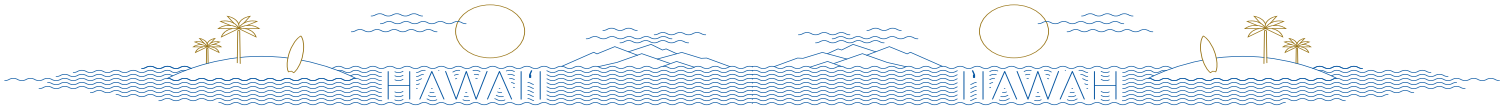
Topics

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



Experiment Design

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



Experiment Terminology (Part 2)

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

Within-subjects, Between-subjects

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

Within-subjects

Participant	Test Condition		
1	A	B	C
2	A	B	C

Between-subjects

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

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!)
  - Comparisons using participant feedback possible
- Disadvantage
  - Order effects (i.e., potential interference between conditions)

Between-subjects

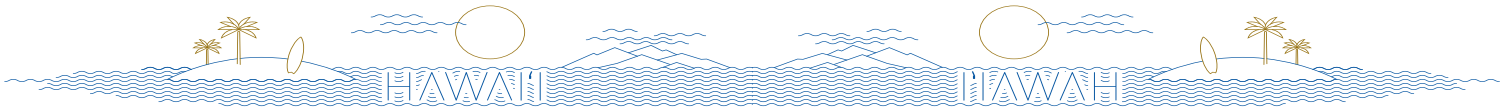
- Disadvantages
  - More participants (harder to recruit, schedule, etc.)
  - More "variation due to participants"
  - Need to balance groups (to ensure they are more or less the same)
  - Comparisons using participant feedback not possible
- Advantage
  - No order effects (i.e., no interference between conditions)

Within-subjects, Between-subjects (2)

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

Counterbalancing

- Needed for within-subjects designs:
  - Participants may benefit from the 1st condition and thereby perform better on the 2nd condition
  - This is a problem (results are misleading)
- To compensate, ***counterbalancing*** is used:
  - Participants are divided into ***groups***, and a different testing order is used for each group
- The testing order is usuallly 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)

Topics

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

Answering Research Questions

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

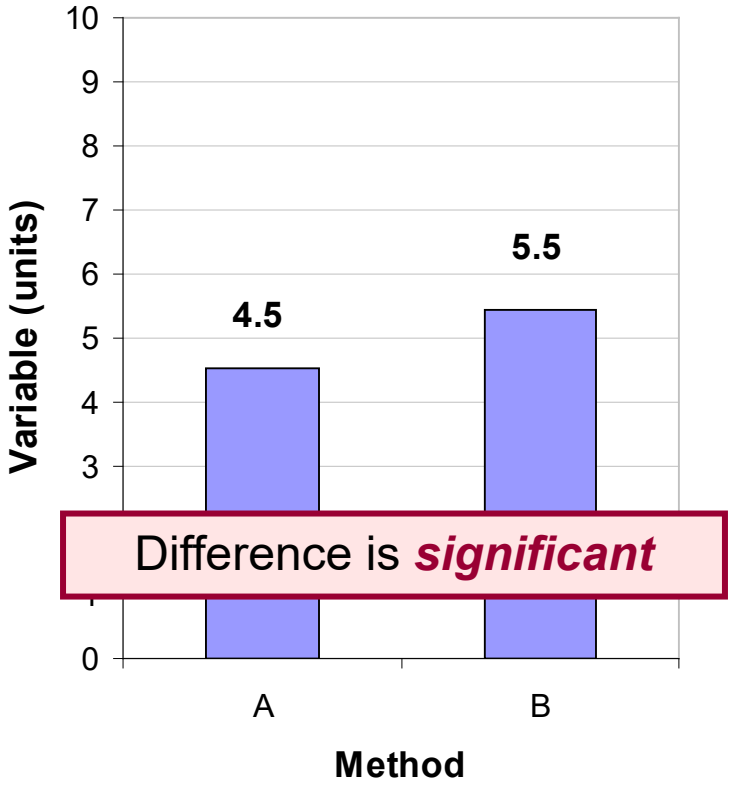
> There is no difference in entry speed between Method A and Method B.

- Assumption of "no difference"
- Research usually 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?

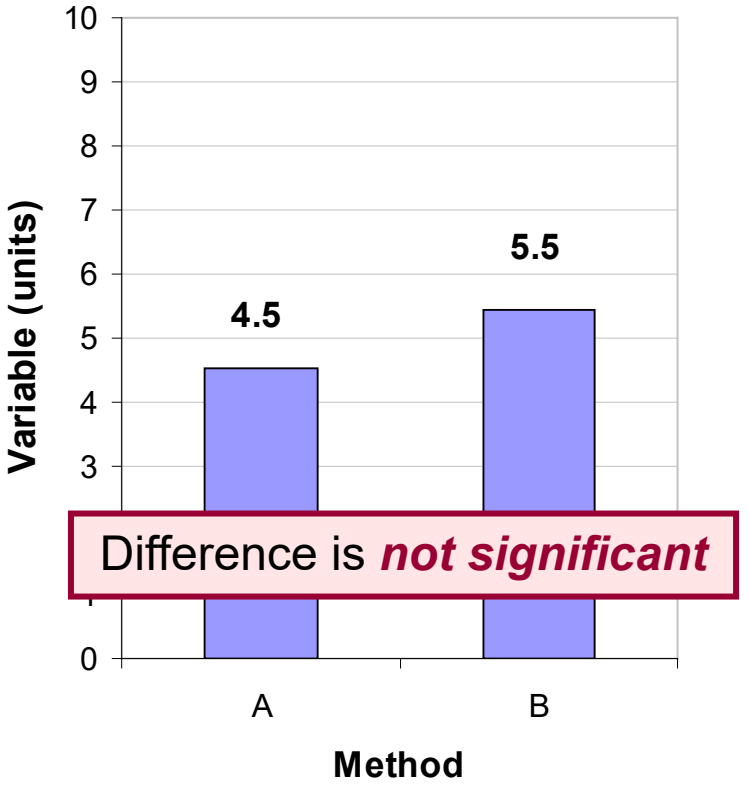
Example #1



Difference is ***significant***

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

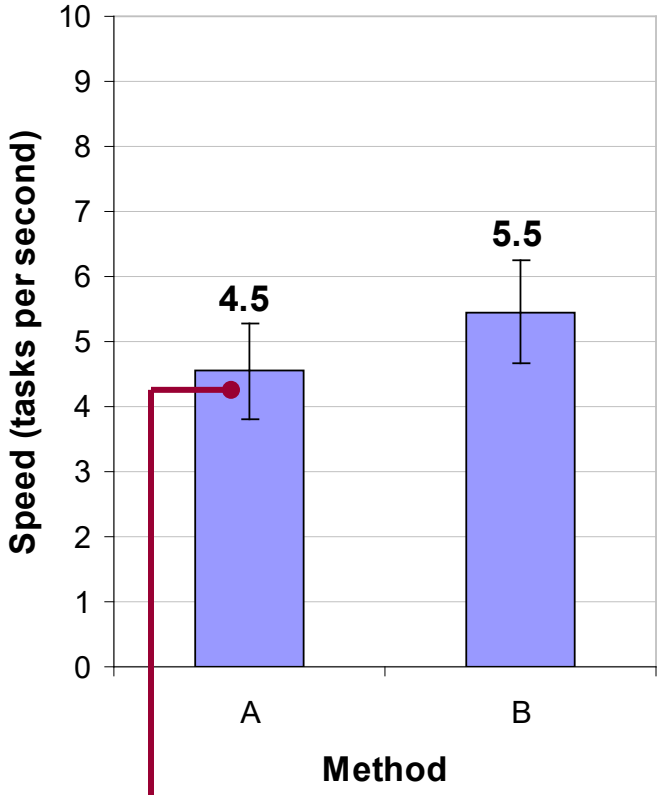
Example #2



Difference is ***not significant***

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

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

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…

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

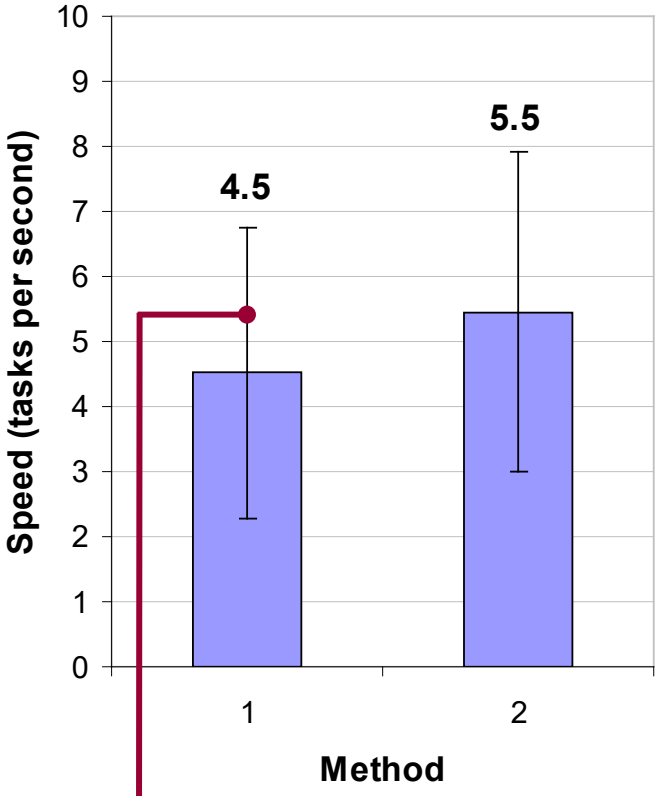
Thresholds for "p"
- .05
- .01
- .005
- .001
- .0005
- .0001

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

Example #2 - Details



Error bars show
±1 standard deviation

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

Example #2 – ANOVA

ANOVA Table for Task Completion Time (s)

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

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

Reported as…

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

Two ways of reporting non-significant effects:

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

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

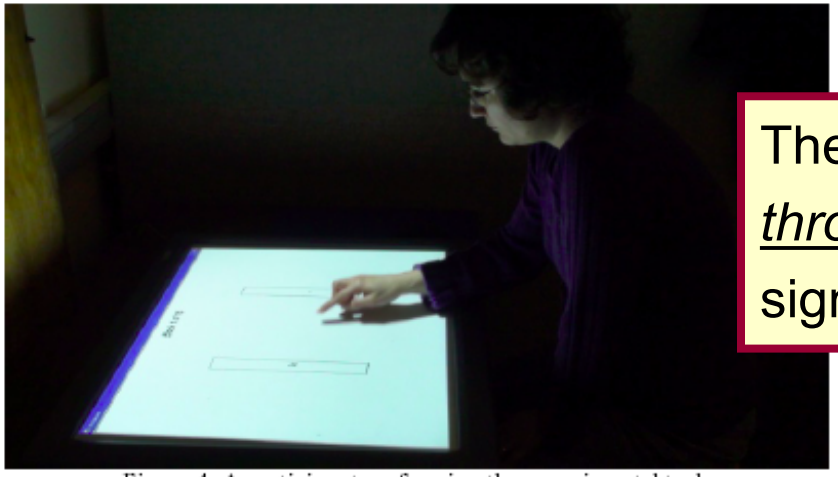


Figure 4. A participant performing the experimental task

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

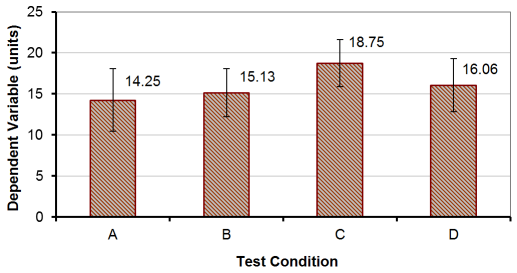
Independent variable:
Input technique

Dependent variable:
Throughput

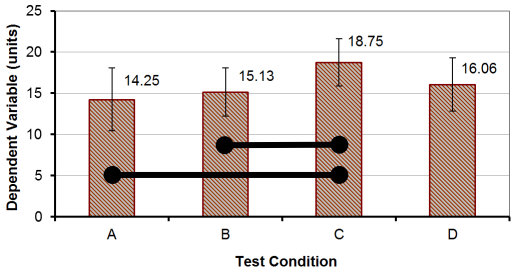
Sasangohar, F., MacKenzie, I. S., & Scott, S. D. (2009). Evaluation of mouse and touch input for a tabletop display using Fitts' reciprocal tapping task. *Proc HFES 2009*, pp. 839-843.

Other Designs and Procedures

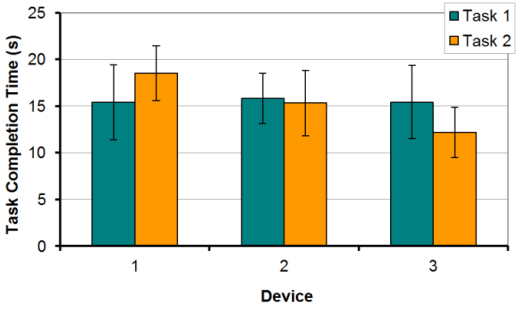
- 1 factor with 4 levels →



- With post hoc tests →

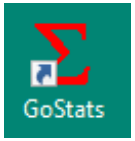


- Two-factor design →



- etc.

ANOVA Demos



- *StatView* (now sold as JMP, http://jmp.com)
  - Commercial statistics package
  - Input data shown on the right
- *GoStats*
  - Java program and its API are freely available via the URL on the last slide
  - Input file (same data as on the right): `anova-ex1.txt`

Example 1 data

Method A	Method B
5.3	5.7
3.6	4.8
5.2	5.1
3.5	4.5
4.6	6.0
4.1	6.8
4.0	6.0
4.8	4.6
5.2	5.5
5.1	5.6

ANOVA Demos (2)

StatView

ANOVA Table for Task Completion Time (s)

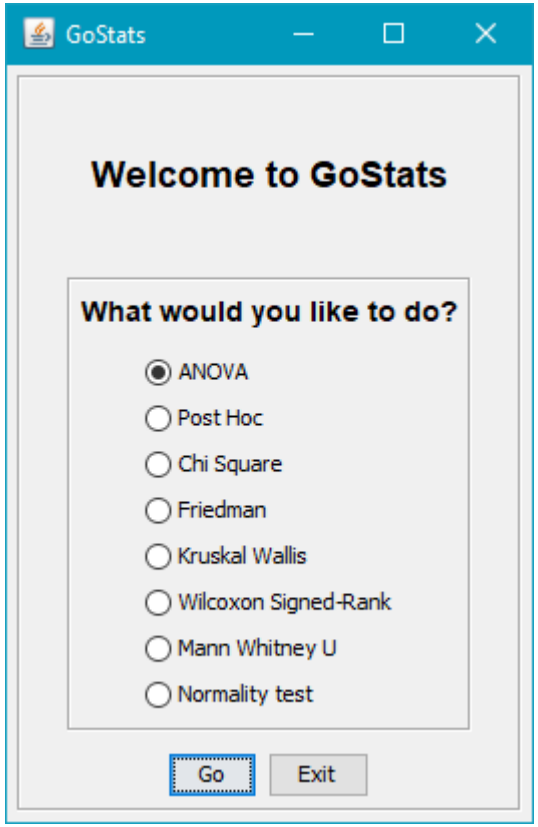
	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.080	.564				
Method	1	4.232	4.232	9.796	.0121	9.796	.804
Method * Subject	9	3.888	.432				

GoStats

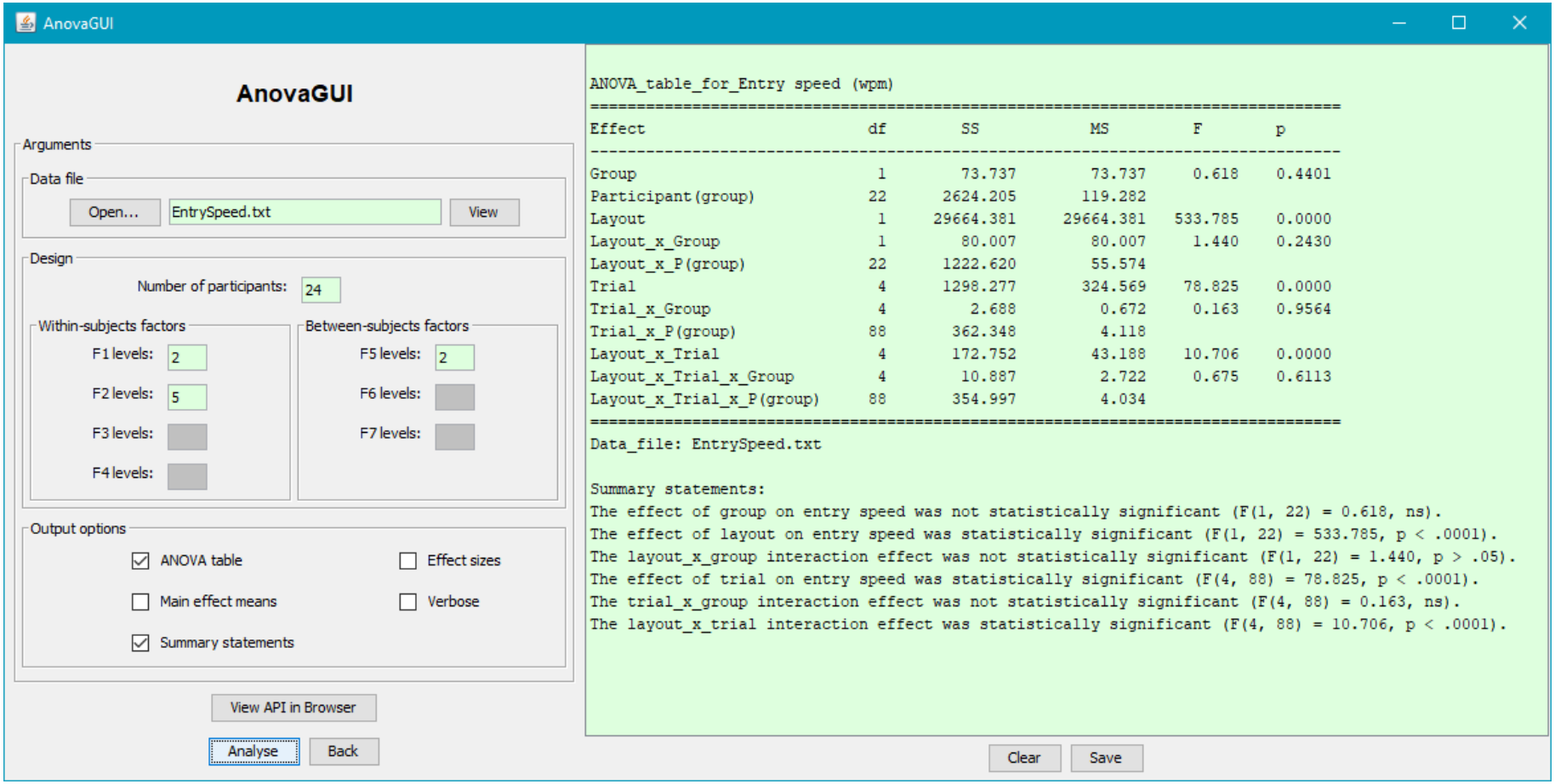
```
ANOVA_table
=============================================================================
Effect                    df        SS              MS          F         p
-----------------------------------------------------------------------------
Participant                9         5.080          0.564
F1                         1         4.232          4.232       9.796     0.0121
F1_x_Par                   9         3.888          0.432
=============================================================================
Data_file: anova-ex1.txt
```

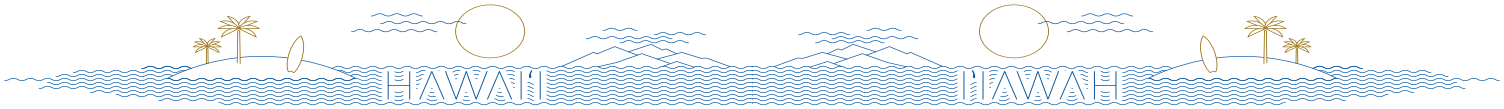
GoStats (1)

- Available as free download via the URL on the last slide
- Includes tools for:
  - ANOVA
  - Post Hoc
  - Other statistical tests conducted in HCI research
- Also includes an API with example experiments and input files



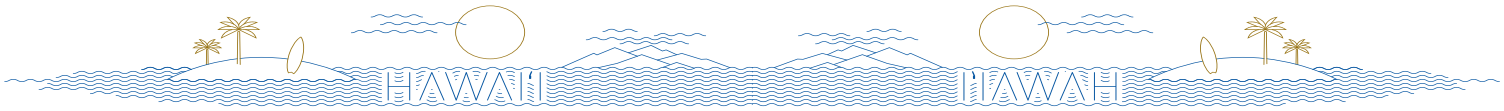
GoStats (2)


AnovaGUI
AnovaGUI
Arguments
Data file
Open...
EntrySpeed.txt
View
Design
Number of participants:
24
Within-subjects factors
F1 levels:
2
F2 levels:
5
F3 levels:
F4 levels:
Between-subjects factors
F5 levels:
2
F6 levels:
F7 levels:
Output options
ANOVA table
Effect sizes
Main effect means
Verbose
Summary statements
View API in Browser
Analyse
Back
ANOVA_table_for_Entry speed (wpm)
Effect df SS MS F p
Group 1 73.737 73.737 0.618 0.4401
Participant(group) 22 2624.205 119.282
Layout 1 29664.381 29664.381 533.785 0.0000
Layout_x_Group 1 80.007 80.007 1.440 0.2430
Layout_x_P(group) 22 1222.620 55.574
Trial 4 1298.277 324.569 78.825 0.0000
Trial_x_Group 4 2.688 0.672 0.163 0.9564
Trial_x_P(group) 88 362.348 4.118
Layout_x_Trial 4 172.752 43.188 10.706 0.0000
Layout_x_Trial_x_Group 4 10.887 2.722 0.675 0.6113
Layout_x_Trial_x_P(group) 88 354.997 4.034
Data_file: EntrySpeed.txt
Summary statements:
The effect of group on entry speed was not statistically significant (F(1, 22) = 0.618, ns).
The effect of layout on entry speed was statistically significant (F(1, 22) = 533.785, p < .0001).
The layout_x_group interaction effect was not statistically significant (F(1, 22) = 1.440, p > .05).
The effect of trial on entry speed was statistically significant (F(4, 88) = 78.825, p < .0001).
The trial_x_group interaction effect was not statistically significant (F(4, 88) = 0.163, ns).
The layout_x_trial interaction effect was statistically significant (F(4, 88) = 10.706, p < .0001).
Clear
Save



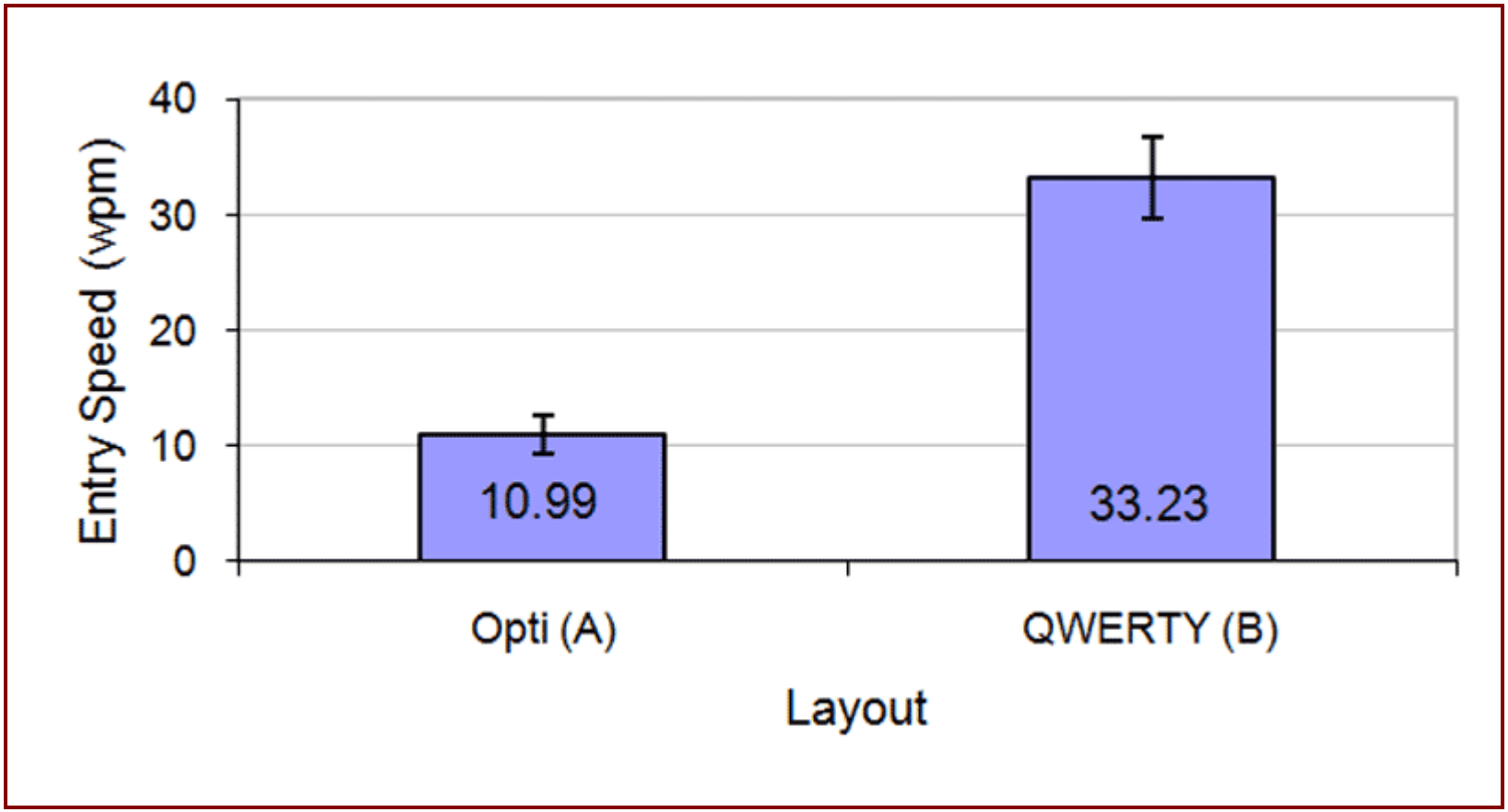
Group Participation Results

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

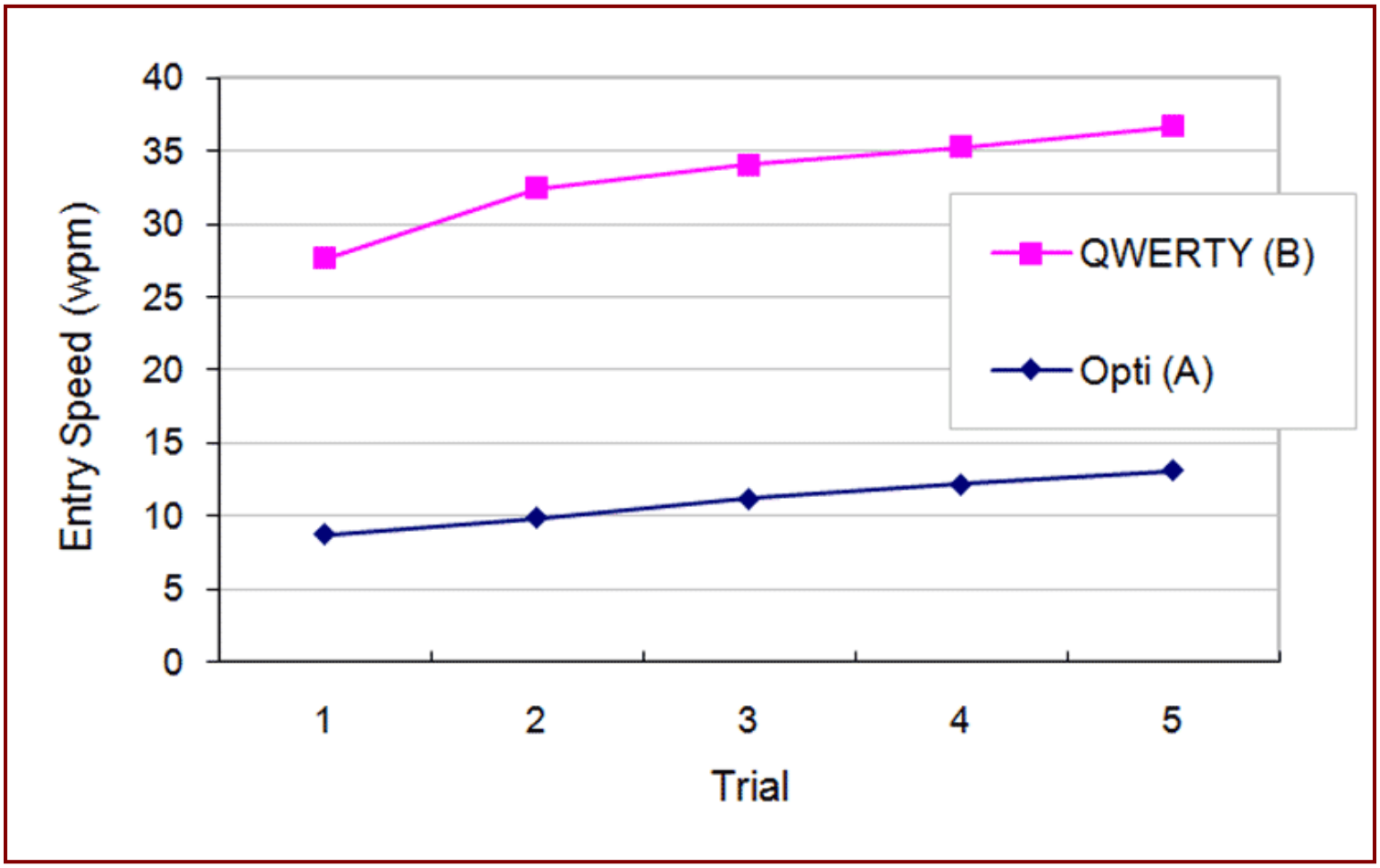


Entry Time (seconds)												
Participant	Initials	Opti (A)					QWERTY (B)					Group
		1	2	3	4	5	1	2	3	4	5	
P1	al	92.0	94.0	84.0	68.0	93.0	23.0	19.0	17.0	17.0	15.0	1
P2	ig	65.0	63.0	55.0	49.0	41.0	18.0	15.0	14.0	14.0	13.0	1
P3	ma	54.0	44.0	38.0	38.0	32.0	19.0	17.0	17.0	15.0	19.0	1
P4	kw	65.0	71.0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

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	



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).

Created using GoStats

ANOVA_table_for_Entry speed (wpm)

Effect	df	SS	MS	F	p
Group	1	73.737	73.737	0.618	0.4401
Participant(group)	22	2624.205	119.282		
Layout	1	29664.381	29664.381	533.785	0.0000
Layout_x_Group	1	80.007	80.007	1.440	0.2430
Layout_x_P(group)	22	1222.620	55.574		
Trial	4	1298.277	324.569	78.825	0.0000
Trial_x_Group	4	2.688	0.672	0.163	0.9564
Trial_x_P(group)	88	362.348	4.118		
Layout_x_Trial	4	172.752	43.188	10.706	0.0000
Layout_x_Trial_x_Group	4	10.887	2.722	0.675	0.6113
Layout_x_Trial_x_P(group)	88	354.997	4.034		

Data_file: EntrySpeed.txt

- 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

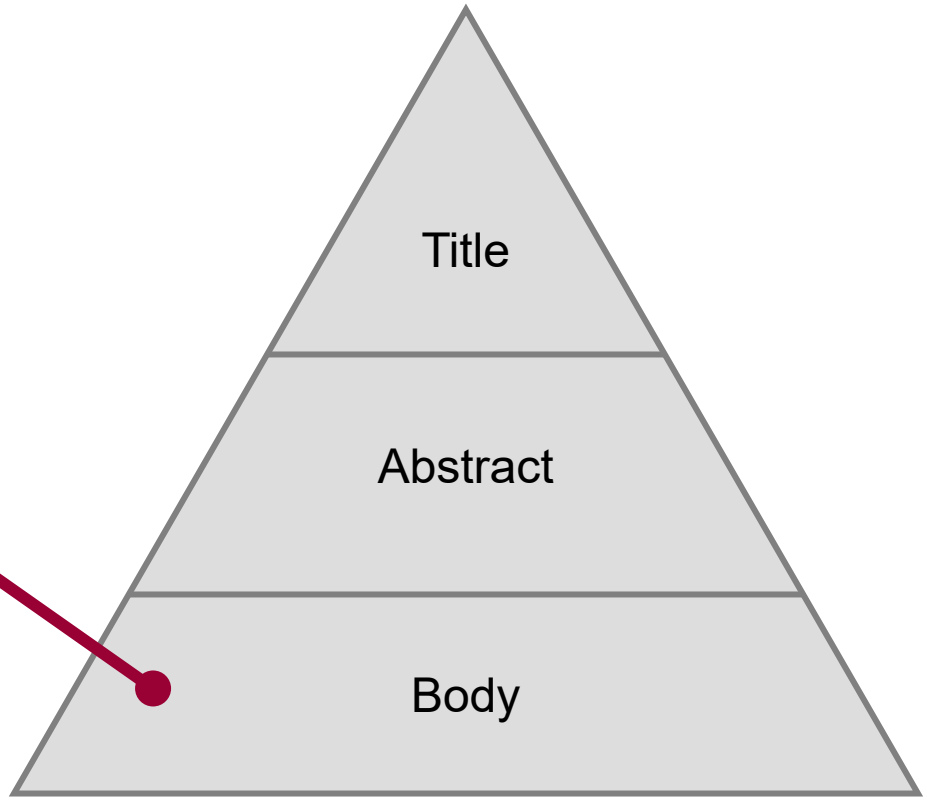
Topics

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

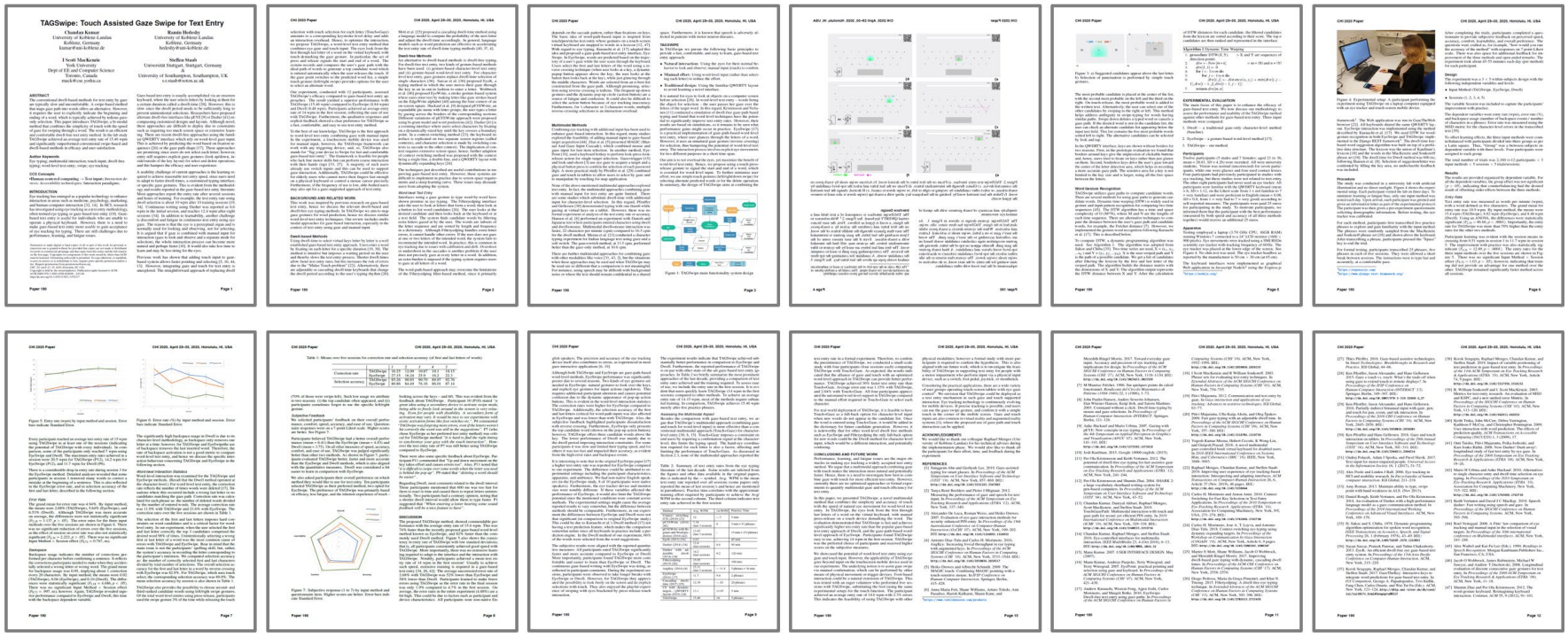
Research Paper

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

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

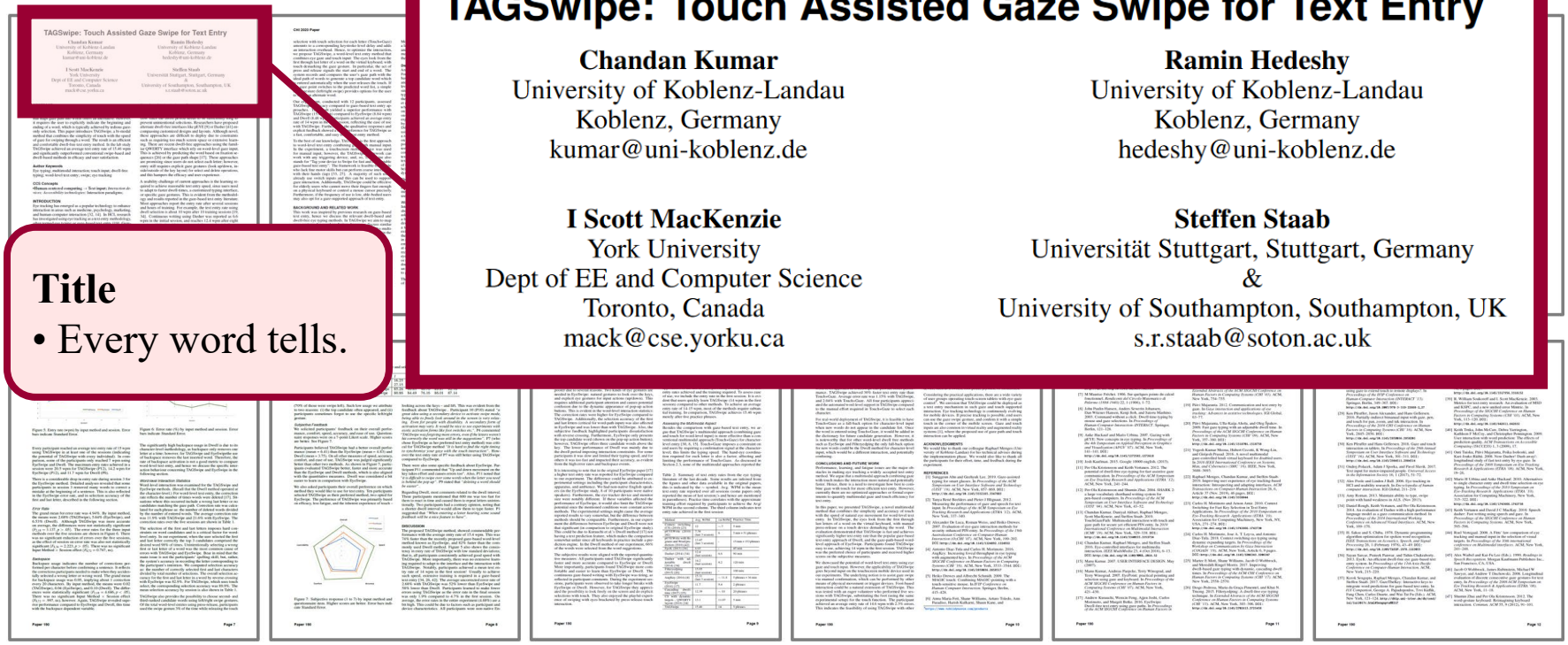


Example Publication†



† Kumar, C., Hedeshy, R., MacKenzie, I. S., & Staab, S. (2020). TAGSwipe: Touch assisted gaze swipe for text entry. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing System – CHI 2020*, pp. 190:1-190:12. New York, ACM. doi:10.1145/3313831.3376317.

Title, Author(s), Affiliation(s)



TAGSwipe: Touch Assisted Gaze Swipe for Text Entry

Chandan Kumar
University of Koblenz-Landau
Koblenz, Germany
kumar@uni-koblenz.de

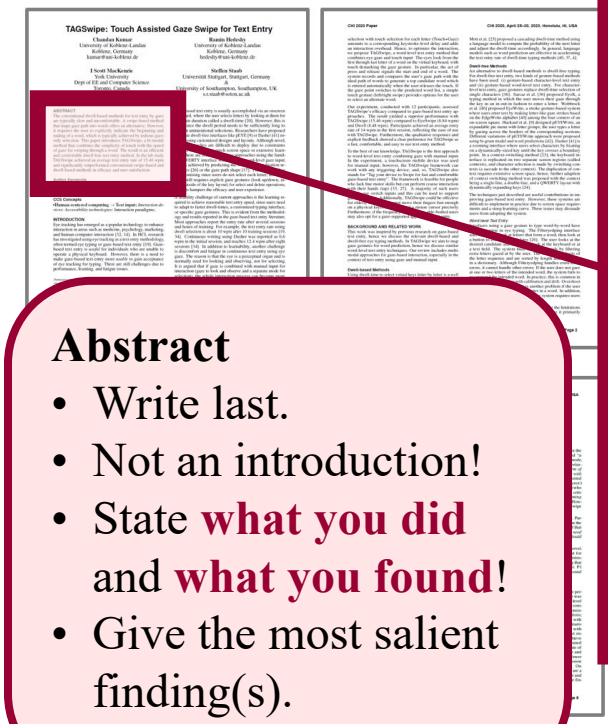
Ramin Hedeshy
University of Koblenz-Landau
Koblenz, Germany
hedeshy@uni-koblenz.de

I Scott MacKenzie
York University
Dept of EE and Computer Science
Toronto, Canada
mack@cse.yorku.ca

Steffen Staab
Universität Stuttgart, Stuttgart, Germany
&
University of Southampton, Southampton, UK
s.r.staab@soton.ac.uk

Title
- Every word tells.

Abstract



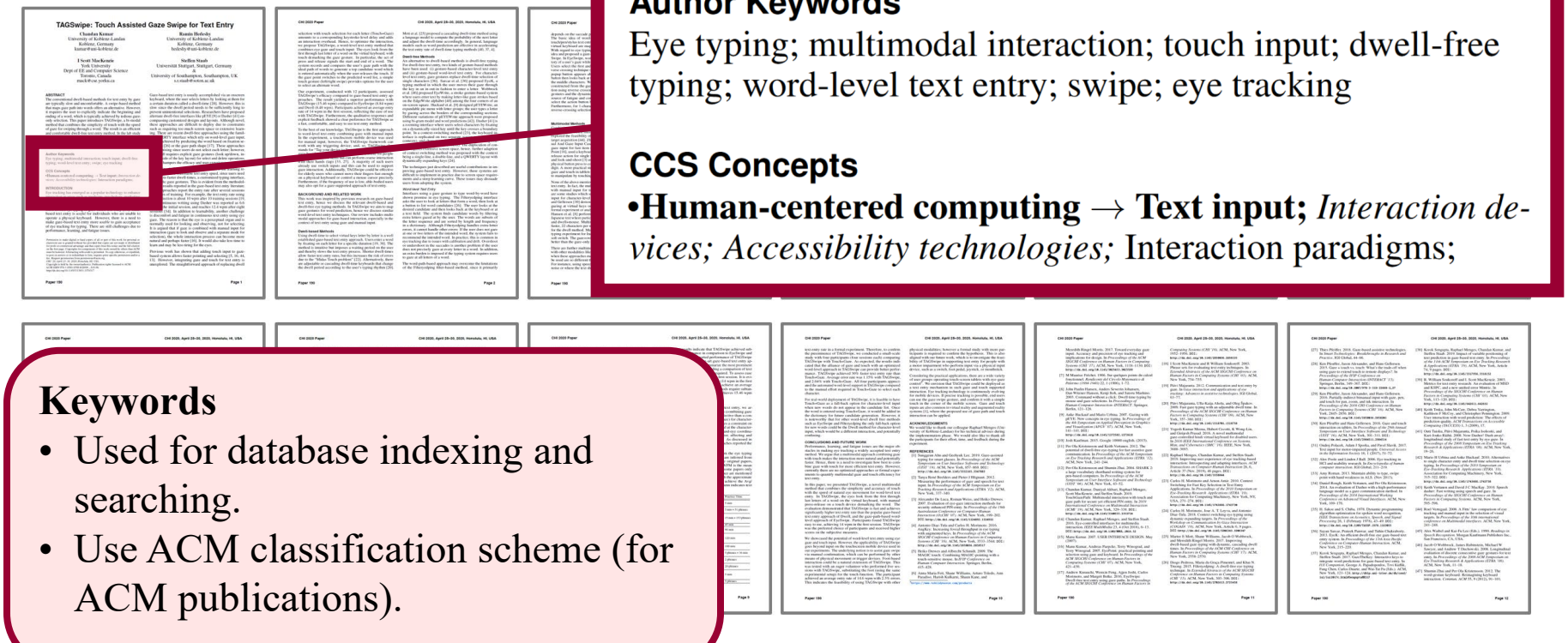
Abstract

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

ABSTRACT

The conventional dwell-based methods for text entry by gaze are typically slow and uncomfortable. A swipe-based method that maps gaze path into words offers an alternative. However, it requires the user to explicitly indicate the beginning and ending of a word, which is typically achieved by tedious gaze-only selection. This paper introduces TAGSwipe, a bi-modal method that combines the simplicity of touch with the speed of gaze for swiping through a word. The result is an efficient and comfortable dwell-free text entry method. In the lab study TAGSwipe achieved an average text entry rate of 15.46 wpm and significantly outperformed conventional swipe-based and dwell-based methods in efficacy and user satisfaction.

Keywords



Author Keywords

Eye typing; multimodal interaction; touch input; dwell-free typing; word-level text entry; swipe; eye tracking

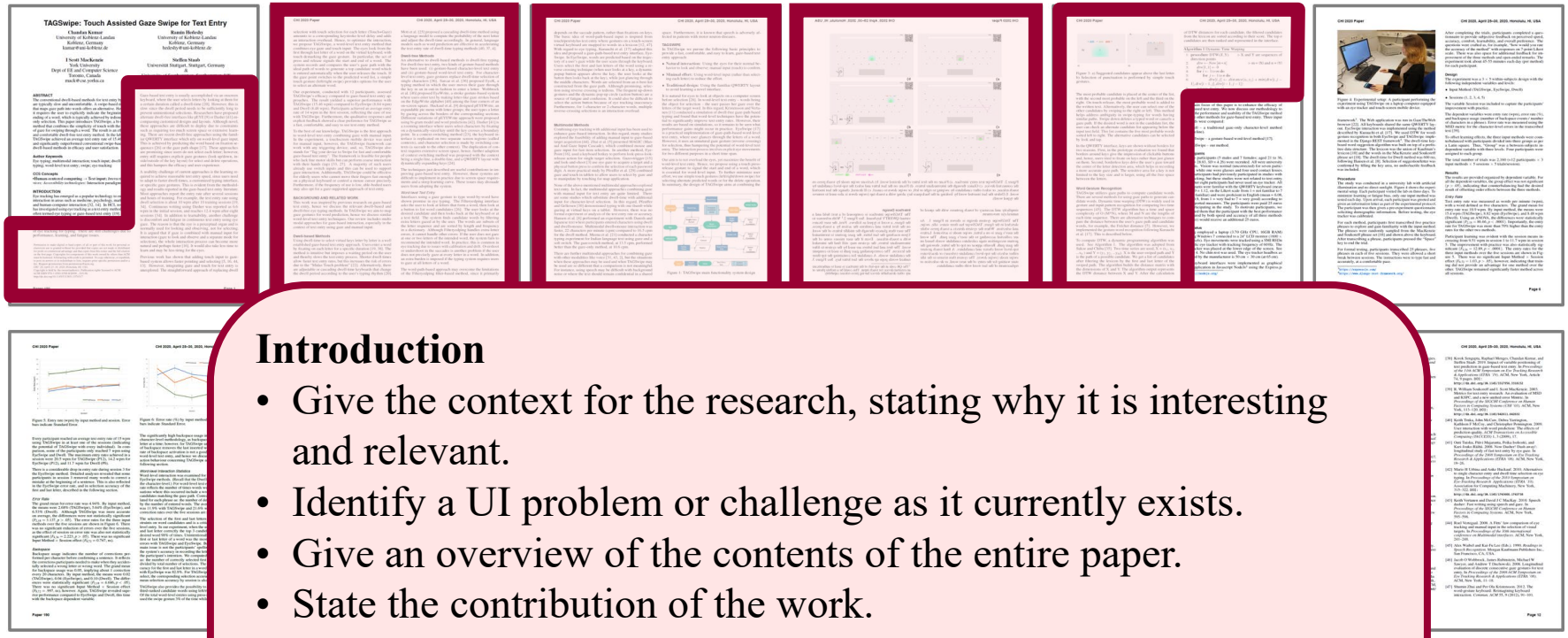
CCS Concepts

•**Human-centered computing** → **Text input;** *Interaction devices; Accessibility technologies;* Interaction paradigms;

Keywords

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

Introduction

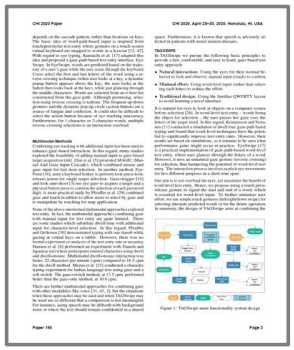


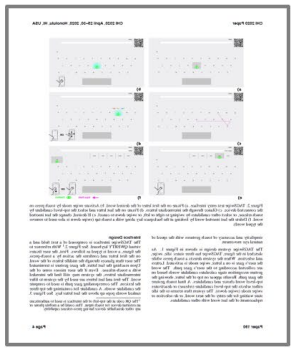
Introduction

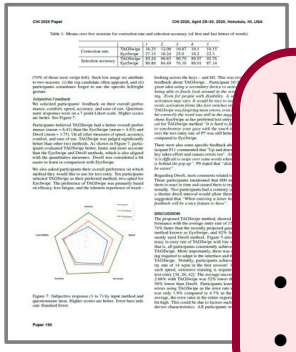
- Give the context for the research, stating why it is interesting and relevant.
- Identify a UI problem or challenge as it currently exists.
- Give an overview of the contents of the entire paper.
- State the contribution of the work.
- Identify, describe, cite related work.
- Describe and justify your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

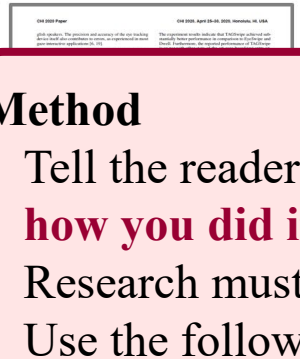
Method

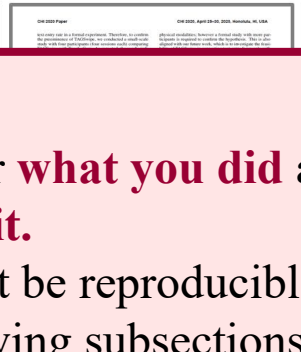

TAGSwipe: Touch Assisted Gaze Swipe for Text Entry











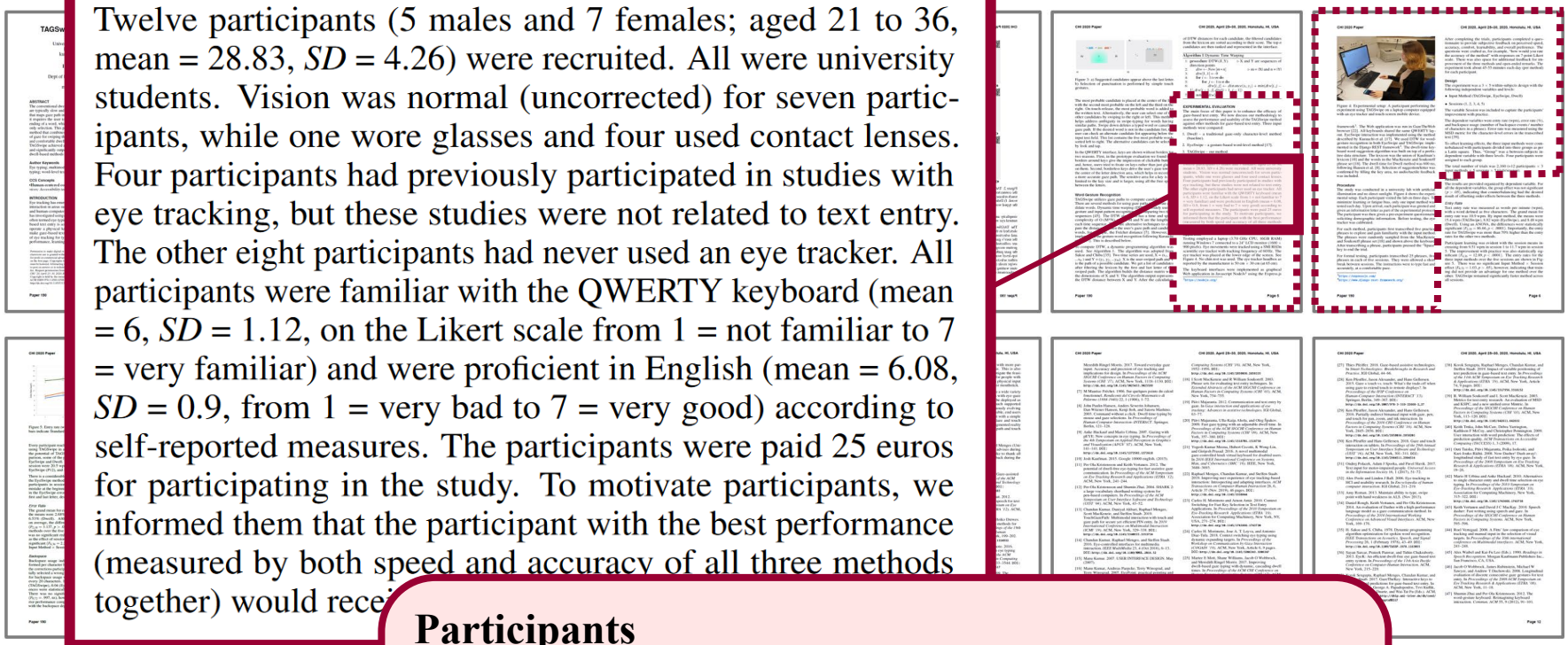
Method

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

Method - Participants

Participants

Twelve participants (5 males and 7 females; aged 21 to 36, mean = 28.83, *SD* = 4.26) were recruited. All were university students. Vision was normal (uncorrected) for seven participants, while one wore glasses and four used contact lenses. Four participants had previously participated in studies with eye tracking, but these studies were not related to text entry. The other eight participants had never used an eye tracker. All participants were familiar with the QWERTY keyboard (mean = 6, *SD* = 1.12, on the Likert scale from 1 = not familiar to 7 = very familiar) and were proficient in English (mean = 6.08, *SD* = 0.9, from 1 = very bad to 7 = very good) according to self-reported measures. The participants were paid 25 euros for participating in the study. To motivate participants, we informed them that the participant with the best performance (measured by both speed and accuracy of all three methods together) would rece



Participants

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

Method - Apparatus

Apparatus

Testing employed a laptop (3.70 GHz CPU, 16GB RAM) running Windows 7 connected to a 24" LCD monitor (1600 × 900 pixels). Eye movements were tracked using a SMI REDn scientific eye tracker with tracking frequency of 60 Hz. The eye tracker was placed at the lower edge of the screen. See Figure 4. No chin rest was used. The eye tracker headbox as reported by the manufacturer is 50 cm × 30 cm (at 65 cm).

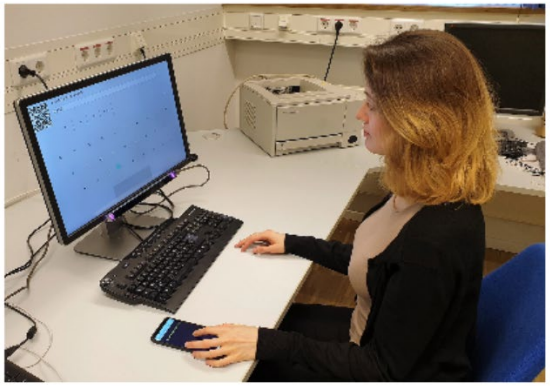


Figure 4: Experimental setup: A participant performing the experiment using TAGSwipe on a laptop computer equipped with an eye tracker and touch-screen mobile device.

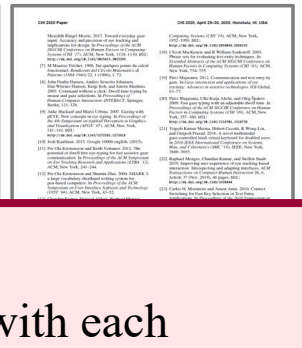
Apparatus

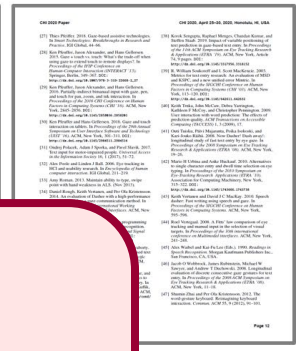
- Describe the hardware and software.
- Use screen snaps or photos, if helpful.

phrase set [18]. The dwell-time for Dwell method was 600 ms, following Hansen et al. [8]. Selection of suggestion/letter was confirmed by filling the key area, no audio/tactile feedback was included.

Method - Procedure


TAGSwipe: Touch Assisted Gaze Swipe for Text Entry





Procedure

The study was conducted in a university lab with artificial illumination and no direct sunlight. Figure 4 shows the experimental setup. Each participant visited the lab on three days. To minimize learning or fatigue bias, only one input method was tested each day. Upon arrival, each participant was greeted and given an information letter as part of the experimental protocol. The participant was then given a pre-experiment questionnaire soliciting demographic information. Before testing, the eye tracker was calibrated.

For each method, participants first transcribed five practice phrases to explore and gain familiarity with the input method. The phrases were randomly sampled from the MacKenzie and Soukoreff phrase set [18] and shown above the keyboard. After transcribing a phrase, participants pressed the "Space" key to end the trial.

For formal testing, participants transcribed 25 phrases, five phrases in each of five sessions. They were allowed a short break between sessions. The instructions were to type fast and accurately, at a

After comp
tionnaire t
accuracy,
questions
the accurac
scale. The
provement
experimen
for each pa

Procedure

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

Method - Design


TAGSwipe: Touch Assisted Gaze Swipe for Text Entry

Design

The experiment was a 3 × 5 within-subjects design with the following independent variables and levels:

- Input Method (TAGSwipe, EyeSwipe, Dwell)
- Sessions (1, 2, 3, 4, 5)

The variable Session was included to capture the participants' improvement with practice.

The dependent variables were entry rate (wpm), error rate (%), and backspace usage (number of backspace events / number of characters in a phrase). Error rate was measured using the MSD
text

To c
terb
a L
dep
assi

The
input methods × 5 sessions × 5 trials/session).

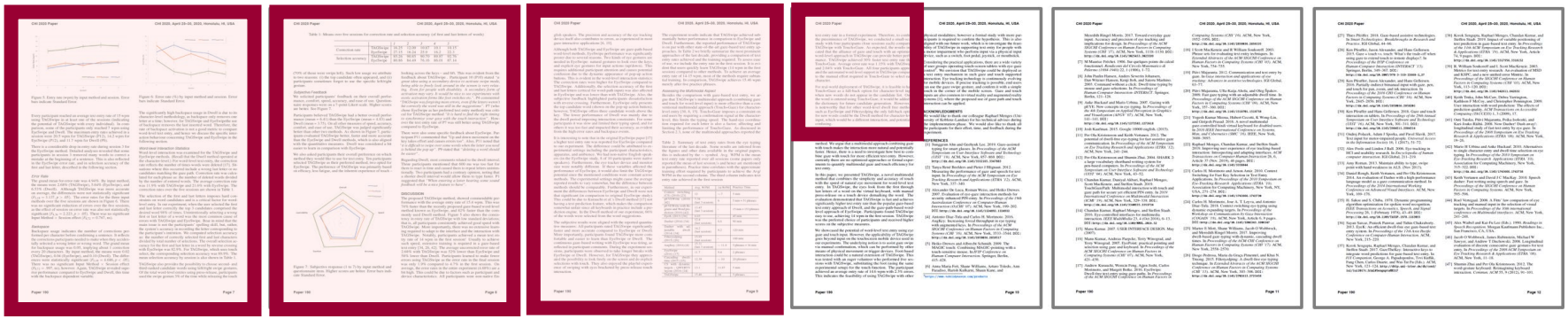
Design

- 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.

Results and Discussion (1)

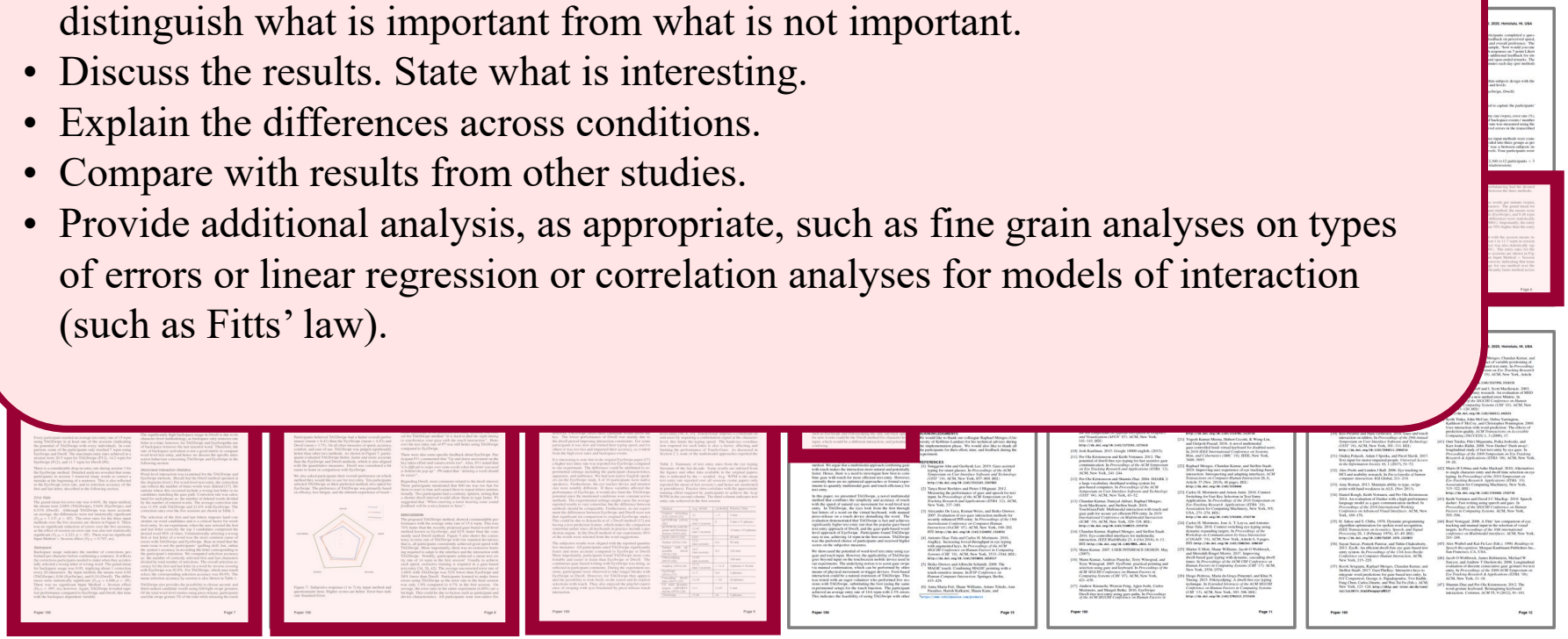
Results and Discussion

- 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.



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).

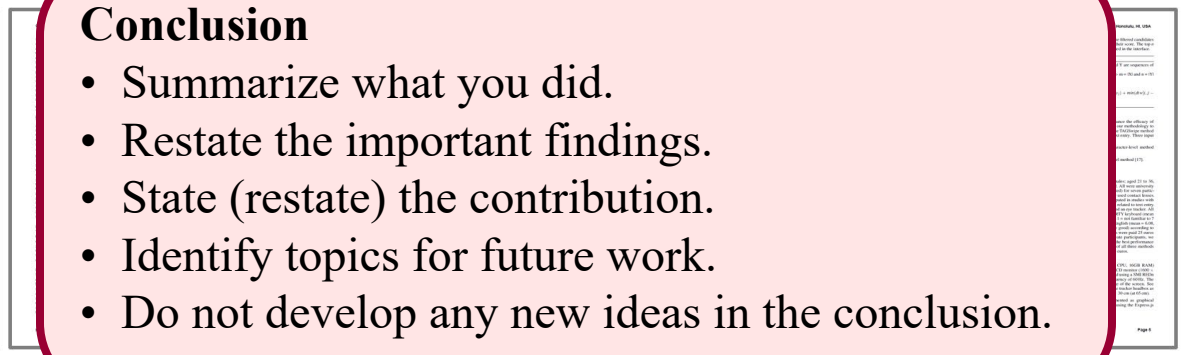


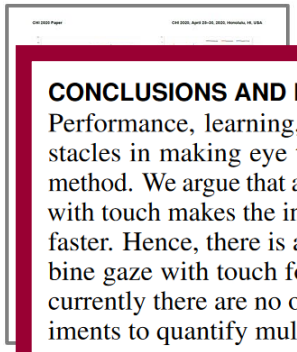
Conclusion

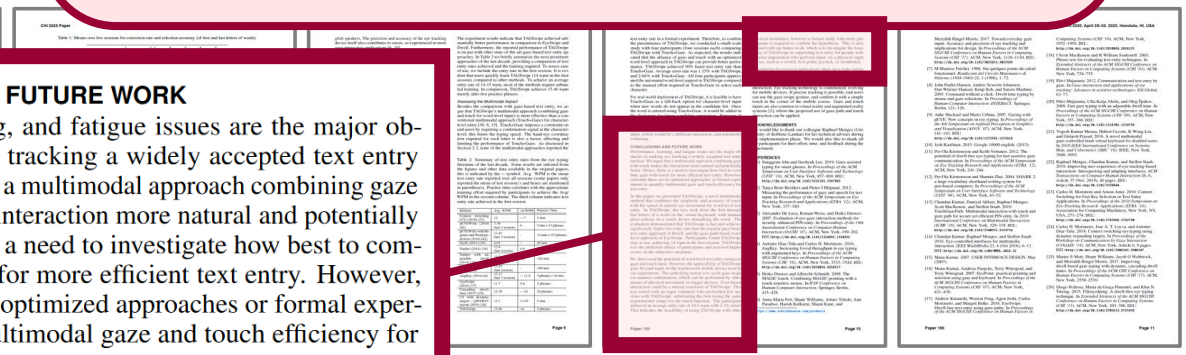
Conclusion

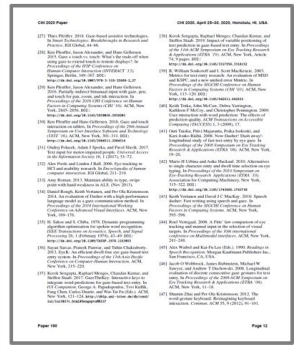
- Summarize what you did.
- Restate the important findings.
- State (restate) the contribution.
- Identify topics for future work.
- Do not develop any new ideas in the conclusion.


TAGSwipe: Touch Assisted Gaze Swipe for Text Entry









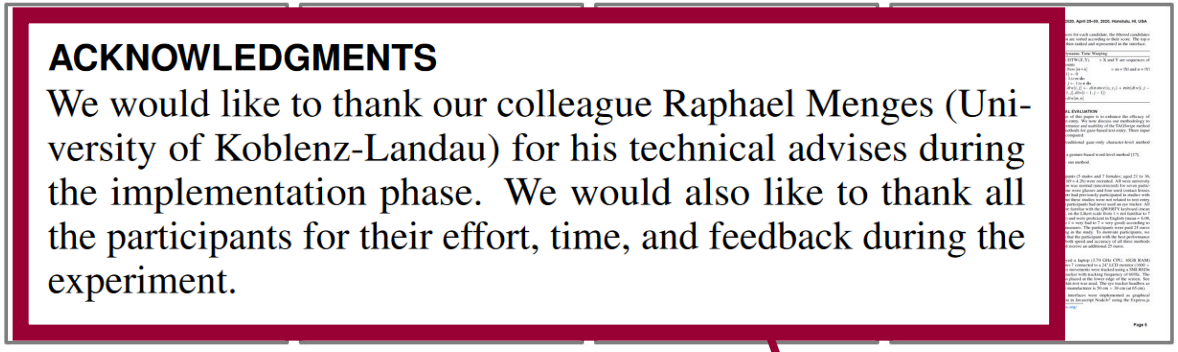
CONCLUSIONS AND FUTURE WORK

Performance, learning, and fatigue issues are the major obstacles in making eye tracking a widely accepted text entry method. We argue that a multimodal approach combining gaze with touch makes the interaction more natural and potentially faster. Hence, there is a need to investigate how best to combine gaze with touch for more efficient text entry. However, currently there are no optimized approaches or formal experiments to quantify multimodal gaze and touch efficiency for text entry.

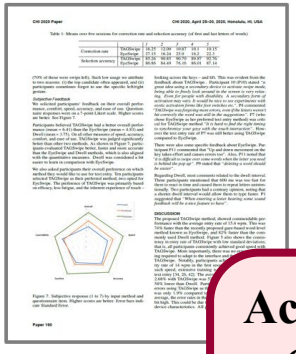
In this paper, we presented TAGSwipe, a novel multimodal method that combines the simplicity and accuracy of touch with the speed of natural eye movement for word-level text entry. In TAGSwipe, the eyes look from the first through last letters of a word on the virtual keyboard, with manual press-release on a touch device demarking the word. The evaluation demonstrated that TAGSwipe is fast and achieves significantly higher text entry rate than the popular gaze-based

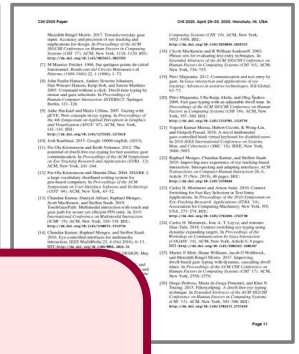
Acknowledgment

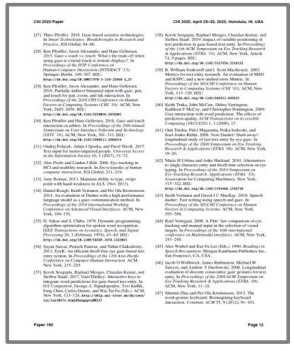

TAGSwipe: Touch Assisted Gaze Swipe for Text Entry


ACKNOWLEDGMENTS

We would like to thank our colleague Raphael Menges (University of Koblenz-Landau) for his technical advises during the implementation phase. We would also like to thank all the participants for their effort, time, and feedback during the experiment.



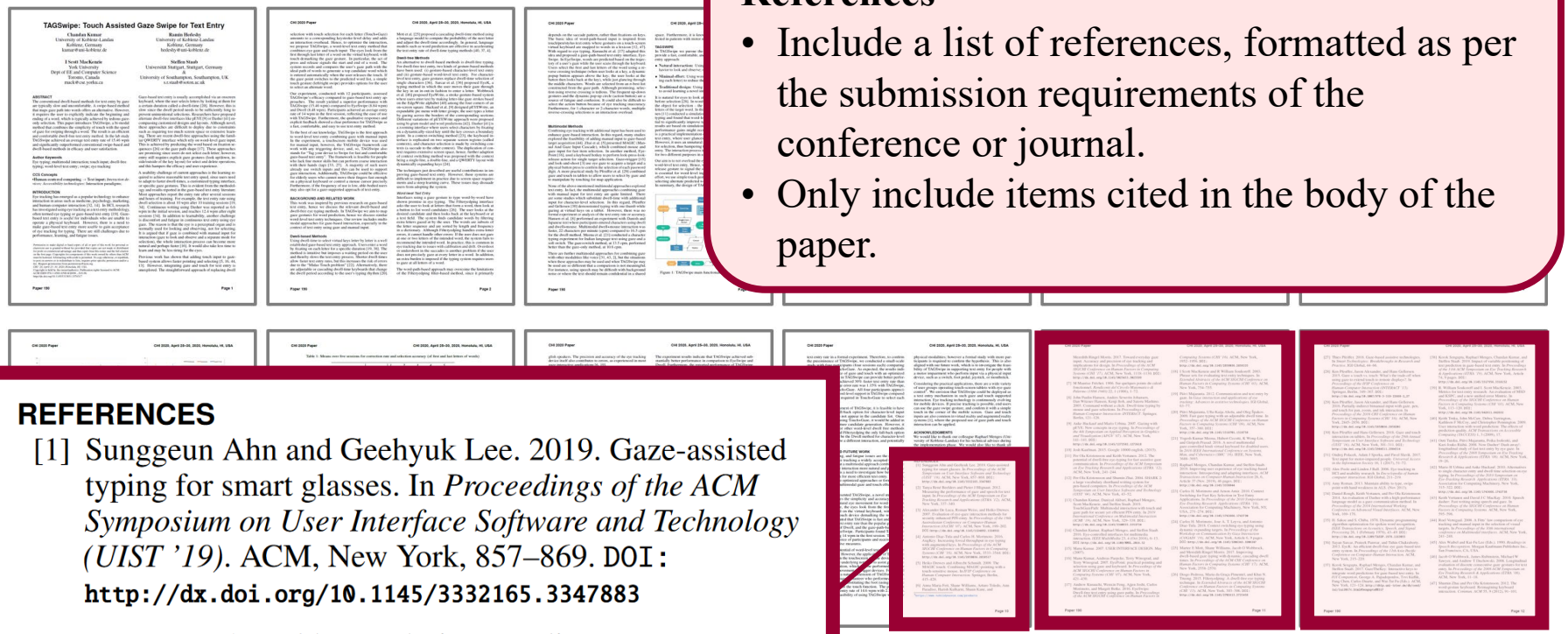




Acknowledgment

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

References



References

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

REFERENCES

[1] Sunggeun Ahn and Geehyuk Lee. 2019. Gaze-assisted typing for smart glasses. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '19)*. ACM, New York, 857–869. DOI: **http://dx.doi.org/10.1145/3332165.3347883**

[2] Tanya René Beelders and Pieter J Blignaut. 2012. Measuring the performance of gaze and speech for text input. In *Proceedings of the ACM Symposium on Eye Tracking Research and Applications (ETRA '12)*. ACM, New York, 337–340.

Summary

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

Thank you

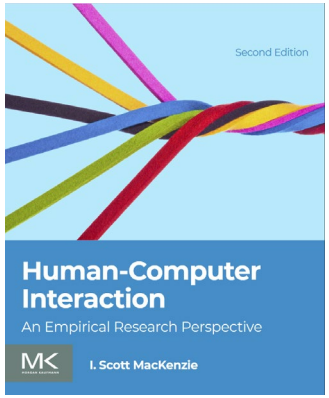
Course slides, etc., are at…

https://www.yorku.ca/mack/CHI2024/

Course feedback survey…

https://www.surveymonkey.com/r/HZFVZ66

For the complete story, see Scott's book:



http://www.yorku.ca/mack/HCIbook2e