Empirical Research Methods in Human-Computer Interaction

I. Scott MacKenzie
York University

Part I – Briefly
What is Empirical Research?

• Empirical research is…
  • observation-based investigation seeking to discover and interpret facts, theories, or laws (relating to 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
  • Develop or explore models that describe or predict behaviour (of humans interacting with computers)
How do we do Empirical Research?

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

† Scientific method - a body of techniques for investigating phenomena and acquiring new knowledge, as well as for correcting and integrating previous knowledge. It is based on gathering observable, empirical, measurable evidence, subject to the principles of reasoning. (wikipedia)

Part II – The Details
(with an HCl context)
Themes

- Observe and measure
- Research questions
- User studies – group participation
- User studies – terminology
- User studies – step by step summary
- Parts of a research paper

When we do Research, we…

- Observe
- Measure
- Describe
- Compare
- Infer
- Relate
- Predict
- etc.
Empirical
When we do Research, we…

• Observe … human behaviour and response
• Measure … using numbers
• Describe … using numbers
• Compare … using numbers
• Infer … using numbers
• Relate … using numbers
• Predict … using numbers
• etc.

Empirical - capable of being verified or disproved by observation or experiment (Websters dictionary)

So, what is non-empirical research?

Non-Empirical Research

• Non-empirical research (aka qualitative research) is generally concerned with the reasons underlying human behaviour (i.e., the why or how, as opposed to the what, where, or when)
• Tends to focus on human…
  • thought, feeling, attitude, emotion, reflection, sentiment, opinion, mood, outlook, manner, approach, strategy, etc.
  • These human qualities are not directly observable or measurable and, therefore, necessitate a different method of inquiry (e.g., case studies, focus groups, cultural probes, personae, etc.)
• But see… (click here)
Observe

• Observations are gathered…
  • Manually
    • Human observers using log sheets, notebooks, questionnaires, etc.
  • Automatically
    • Sensors, switches, cameras, etc.
    • Computer + software to log events + timestamps

Measure

• A measurement is a recorded observation
• An empirical measurement is a number

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

Kelvin, 1883
Scales of Measurement

- **Nominal**
  - Nominal – arbitrary assignment of a code to an attribute, e.g.,
    - 1 = male, 2 = female

- **Ordinal**
  - Ordinal – rank, e.g.,
    - 1st, 2nd, 3rd, …

- **Interval**
  - Interval – equal distance between units, but no absolute zero point, e.g.,
    - 20°C, 30°C, 40°C, …

- **Ratio**
  - Ratio – absolute zero point, therefore ratios are meaningful, e.g.,
    - 20 wpm, 40 wpm, 60 wpm

Use ratio measurements where possible.

Ratio Measurements

- Preferred scale of measurement
- With ratio measurements summaries and comparisons are strengthened
- Report “counts” as ratios where possible because they facilitate comparisons
- Example – a 10-word phrase was entered in 30 seconds
  - **Bad:** $t = 30$ seconds
  - **Good:** Entry rate $= 10 / 0.5 = 20$ wpm
- Example – two errors were committed while entering a 10-word (50 character) phrase
  - **Bad:** $n = 2$ errors
  - **Good:** Error rate was $2 / 50 = 0.04 = 4\%$
Observe, Measure… Then What?

- Observations and measurements are gathered in a user study (to get “good” data)
- They, we
  - Describe
  - Compare
  - Infer
  - Relate
  - Predict
  - etc.

These are statistical terms.
Fine, but usually our intent is not statistical.
Our intent is founded on simple well-intentioned “research questions”.
Let’s see…

Themes

- Observe and measure
- Research questions
- User studies – group participation
- User studies – terminology
- User studies – step by step summary
- Parts of a research paper
Research Questions

- Consider the following questions about a new or existing UI design or interaction technique:
  - Is it viable?
  - Is it as good as or better than current practice?
  - Which of several design alternatives is best?
  - What are its performance limits and capabilities?
  - What are its strengths and weaknesses?
  - Does it work well for novices, for experts?
  - How much practice is required to become proficient?

Testable Research Questions

- Preceding questions, while unquestionably relevant, are not testable
- Try to re-cast as testable questions (…even though the new question may appear less important)
- Scenario…
  - You have an idea for a new [technique for entering text on a mobile phone] and you think it's pretty good. In fact, you think it is better than [the commonly used multi-tap technique]. You decide to undertake a program of empirical enquiry to evaluate your idea. What are your research questions?
  - Replace […] as appropriate for other research topics
Research Questions (2)

- Very weak (in an empirical sense)
  - Is the new technique any good?
- Weak
  - Is the new technique better than multi-tap?
- Better
  - Is the new technique faster than multi-tap?
- Better still
  - Is the new technique faster than multi-tap within one hour of use?
- Even better
  - If error rates are kept under 2%, is the new technique faster than multi-tap within one hour of use?

A Tradeoff

<table>
<thead>
<tr>
<th>Accuracy of Answer</th>
<th>Breadth of Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Broad</td>
</tr>
<tr>
<td>Low</td>
<td>Narrow</td>
</tr>
</tbody>
</table>

- Internal validity
- External validity

- Is the new technique better than multi-tap?
- If error rates are kept under 2%, is the new technique faster than multi-tap within one hour of use?
Internal Validity

- Definition: The extent to which the effects observed are due to the test conditions (e.g., multitap vs. new)
- Statistically...
  - Differences (in the means) are due to inherent properties of the test conditions
  - Variances are due to participant differences (‘pre-dispositions’)
  - Other potential sources of variance are controlled or exist equally and randomly across the test conditions

External Validity

- Definition: The extent to which results are generalizable to other people and other situations
- Statistically...
  - People
    - The participants are representative of the broader intended population of users
  - Situations
    - Test environment and experimental procedures are representative of real world situations where the interface or technique will be used
Test Environment Example

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

Experimental Procedure Example

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

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

Strive for the Best of Both Worlds

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

- Many user studies in HCI are one-of
  - I.e., a new user interface is designed and a user study is conducted to find strengths and weaknesses
- Much better to do a comparative evaluation
  - I.e., A new user interface is designed and it is compared with an alternative design to determine which is better
- The alternative may be
  - A variation in the new design
  - An established design (perhaps a “baseline condition”)
- More than two interfaces may be compared
- Testable research questions are comparative!
- See the paper in CHI 2006 by Tohidi et al.

Answering Research Questions

- We want to know if the measured performance on a variable (e.g., speed) is different between test conditions, so…
  - We conduct a user study (more on this soon) 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 Empirical Questions (2)

• Four questions:
  1. Is there a difference?
  2. Is the difference large or small?
  3. Is the difference statistically significant (or is it due to chance)?
  4. Is the difference of practical significance?
• Question #1 – obvious (some difference is likely)
• Question #2 – statistics can’t help (Is a 5% difference large or small?)
• Question #3 – statistics can help
• Question #4 – statistics can’t help (Is a 5% difference useful? People resist change!)
• The basic statistical tool for Question #3 is the analysis of variance (anova)

Analysis of Variance

• It is interesting that the test is called an analysis of \textit{variance}, yet it is used to determine if there is a significant difference between the \textit{means}.
• How is this?
Example #1

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

Example #2

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

File: AnovaDemo.xls

Example #1 - Details

Error bars show ±1 standard deviation

Note: SD is the square root of the variance
Example #1 - Anova

ANOVA Table for Speed

<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F-Value</th>
<th>P-Value</th>
<th>Lambda</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>9</td>
<td>5.839</td>
<td>.649</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>1</td>
<td>4.161</td>
<td>4.161</td>
<td>8.443</td>
<td>.0174</td>
<td>8.443</td>
<td>.741</td>
</tr>
<tr>
<td>Method * Subject</td>
<td>9</td>
<td>4.435</td>
<td>.493</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Probability that the difference in the means is due to chance

Reported as…

\[ F_{1,9} = 8.443, p < .05 \]

Thresholds for "p"

• .05
• .01
• .005
• .001
• .0005
• .0001

How to Report an F-statistic

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

• Notice in the parentheses
  • Uppercase for \( F \)
  • Lowercase for \( p \)
  • Italics for \( F \) and \( p \)
  • Space both sides of equal sign
  • Space after comma
  • Space both sides of less than sign
  • Degrees of freedom are subscript, plain, smaller font
  • Three significant figures for \( F \) statistic
  • No zero before the decimal point in the \( p \) statistic (except in Europe)
### Example #2 - Details

**Graph:**
- Speed (tasks per second) vs. Method (1, 2)
- Error bars show ±1 standard deviation

**Table:**

<table>
<thead>
<tr>
<th>Participant</th>
<th>Method A</th>
<th>Method B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.4</td>
<td>6.9</td>
</tr>
<tr>
<td>2</td>
<td>2.7</td>
<td>7.2</td>
</tr>
<tr>
<td>3</td>
<td>3.4</td>
<td>2.6</td>
</tr>
<tr>
<td>4</td>
<td>6.1</td>
<td>1.8</td>
</tr>
<tr>
<td>5</td>
<td>6.4</td>
<td>7.8</td>
</tr>
<tr>
<td>6</td>
<td>5.4</td>
<td>9.2</td>
</tr>
<tr>
<td>7</td>
<td>7.9</td>
<td>4.4</td>
</tr>
<tr>
<td>8</td>
<td>1.2</td>
<td>6.6</td>
</tr>
<tr>
<td>9</td>
<td>3.0</td>
<td>4.8</td>
</tr>
<tr>
<td>10</td>
<td>6.6</td>
<td>3.1</td>
</tr>
</tbody>
</table>

**Mean:**
- A: 4.5
- B: 5.5

**SD:**
- A: 2.23
- B: 2.45

### Example #2 – Anova

**ANOVA Table for Speed**

<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F-Value</th>
<th>P-Value</th>
<th>Lambda</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>9</td>
<td>37.017</td>
<td>4.113</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>1</td>
<td>4.376</td>
<td>4.376</td>
<td>.634</td>
<td>.4462</td>
<td>.34</td>
<td>.107</td>
</tr>
<tr>
<td>Method * Subject</td>
<td>9</td>
<td>62.079</td>
<td>6.898</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Probability that the difference in the means is due to chance**

**Reported as...**

\[ F_{1,9} = 0.634, \text{ ns} \]

**Note:** For non-significant effects, use "ns" if \( F < 1.0 \), or "p > .05" if \( F > 1.0 \).
Anova Demo - StatView†

Files:
AnovaExample1.svd
AnovaExample2.svd

† Now sold as JMP (see http://www.statview.com)

Anova Demo – Anova2 †

Files:
AnovaExample1.txt
AnovaExample2.txt

† This program and its API are available free to attendees of this course. Click here to view API
More Than Two Test Conditions

Two Factors

ANOVA Table for Dependent Variable (units)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>MS</th>
<th>F-Value</th>
<th>P-Value</th>
<th>Lambda</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>15</td>
<td>9.64</td>
<td>15</td>
<td>.001</td>
<td>.001</td>
<td>.999</td>
</tr>
<tr>
<td>Test Condition</td>
<td>3</td>
<td>102.172</td>
<td>66.724</td>
<td>4.564</td>
<td>.047</td>
<td>.486</td>
</tr>
<tr>
<td>Test Condition * Subject</td>
<td>45</td>
<td>551.978</td>
<td>12.297</td>
<td>.25</td>
<td>.000</td>
<td>.000</td>
</tr>
</tbody>
</table>

ANOVA Table for Task Completion Time (sec)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>MS</th>
<th>F-Value</th>
<th>P-Value</th>
<th>Lambda</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>11</td>
<td>34.03</td>
<td>11.34</td>
<td>.001</td>
<td>.001</td>
<td>.999</td>
</tr>
<tr>
<td>Device</td>
<td>11</td>
<td>17.00</td>
<td>1.70</td>
<td>.25</td>
<td>.25</td>
<td>.000</td>
</tr>
<tr>
<td>Task</td>
<td>11</td>
<td>17.00</td>
<td>1.70</td>
<td>.25</td>
<td>.25</td>
<td>.000</td>
</tr>
<tr>
<td>Task * Subject</td>
<td>11</td>
<td>17.00</td>
<td>1.70</td>
<td>.25</td>
<td>.25</td>
<td>.000</td>
</tr>
<tr>
<td>Task * Device</td>
<td>11</td>
<td>17.00</td>
<td>1.70</td>
<td>.25</td>
<td>.25</td>
<td>.000</td>
</tr>
<tr>
<td>Task * Task</td>
<td>11</td>
<td>17.00</td>
<td>1.70</td>
<td>.25</td>
<td>.25</td>
<td>.000</td>
</tr>
<tr>
<td>Task * Task</td>
<td>11</td>
<td>17.00</td>
<td>1.70</td>
<td>.25</td>
<td>.25</td>
<td>.000</td>
</tr>
</tbody>
</table>

Mean: Device 1, Task 1, Subject A: 15.3, Device 2, Task 2: 15.3, Device 3, Task 3: 15.3
Themes

• Observe and measure
• Research questions
• User studies – group participation
• User studies – terminology
• User studies – step by step summary
• 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 three-page handout is distributed to each group (see next slide)
• Read the instructions on the first page and discuss the procedure with your partner
• Your instructor will provide additional information

†This section may be omitted or shortened depending on the time available
Do the Experiment

- The experiment is performed
- This takes about 30 minutes
- After the experiment... break time (lunch?)
- The instructor and an assistant will transcribe the tabulated data into a ready-made spreadsheet
- Results are instantaneous
- After the break... (next slide)
# Results

When we do Empirical Research, we...

- Observe
- Measure
- Describe
- Compare
- Infer
- Relate
- Predict
- etc.

---

### SimpleExperiment-results-TAUCHI-02-2007.xls

<table>
<thead>
<tr>
<th>Participant</th>
<th>Initials</th>
<th>Phone (Method A)</th>
<th>Overly (Method B)</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>SH</td>
<td>21.0</td>
<td>20.0</td>
<td>1</td>
</tr>
<tr>
<td>P2</td>
<td>PV</td>
<td>22.0</td>
<td>22.0</td>
<td>1</td>
</tr>
<tr>
<td>P3</td>
<td>EV</td>
<td>23.0</td>
<td>23.0</td>
<td>1</td>
</tr>
<tr>
<td>P4</td>
<td>MM</td>
<td>24.0</td>
<td>24.0</td>
<td>1</td>
</tr>
<tr>
<td>P5</td>
<td>HV</td>
<td>25.0</td>
<td>25.0</td>
<td>1</td>
</tr>
<tr>
<td>P6</td>
<td>OM</td>
<td>26.0</td>
<td>26.0</td>
<td>1</td>
</tr>
<tr>
<td>P7</td>
<td>TV</td>
<td>27.0</td>
<td>27.0</td>
<td>1</td>
</tr>
<tr>
<td>P8</td>
<td>LM</td>
<td>28.0</td>
<td>28.0</td>
<td>1</td>
</tr>
<tr>
<td>P9</td>
<td>MM</td>
<td>29.0</td>
<td>29.0</td>
<td>1</td>
</tr>
<tr>
<td>P10</td>
<td>SU</td>
<td>30.0</td>
<td>30.0</td>
<td>1</td>
</tr>
<tr>
<td>P11</td>
<td>JR</td>
<td>31.0</td>
<td>31.0</td>
<td>1</td>
</tr>
<tr>
<td>P12</td>
<td>CH</td>
<td>32.0</td>
<td>32.0</td>
<td>2</td>
</tr>
<tr>
<td>P13</td>
<td>PP</td>
<td>33.0</td>
<td>33.0</td>
<td>2</td>
</tr>
<tr>
<td>P14</td>
<td>KT</td>
<td>34.0</td>
<td>34.0</td>
<td>2</td>
</tr>
<tr>
<td>P15</td>
<td>DH</td>
<td>35.0</td>
<td>35.0</td>
<td>2</td>
</tr>
<tr>
<td>P16</td>
<td>BV</td>
<td>36.0</td>
<td>36.0</td>
<td>2</td>
</tr>
<tr>
<td>P17</td>
<td>SK</td>
<td>37.0</td>
<td>37.0</td>
<td>2</td>
</tr>
<tr>
<td>P18</td>
<td>RH</td>
<td>38.0</td>
<td>38.0</td>
<td>2</td>
</tr>
</tbody>
</table>

---

**Observe**

**Measure**
There was a significant effect of keyboard layout on entry speed ($F_{1,47} = 131.2, p < .0001$).
Prediction Equation from a Longitudinal Study

- For Phone: $y = 23.622x^{0.2}$, $R^2 = 0.9957$
- For Qwerty: $y = 16.825x^{0.196}$, $R^2 = 0.9831$

Click here to view paper.
Themes

- Observe and measure
- Research questions
- User studies – group participation
- User studies – terminology
- User studies – step by step summary
- Parts of a research paper

Experiment Design

- Experiment design is the process of deciding which variables to use, what tasks and procedure to use, how many participants to use and how to solicit them, and so on
- Let’s work on the terminology…
Experiment Design - Terminology

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

Participant

• The people participating in an experiment are referred to as participants
• Previously the term subjects was used, but it is no longer in vogue
• When referring specifically to the experiment, use the term participants (e.g., “all participants exhibited a high error rate…”)
• General comments on the problem or conclusions drawn 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
How Many Participants?

• The Answer: It depends!
• Too many:
  • Results are statistically significant, even where the differences are miniscule and of no practical relevance
• Too few:
  • Results are not statistically significant (because of the small sample size), even though there maybe a significant difference in the test conditions (“significant” in the practical sense)
• Guideline:
  • Use approximately the same number of participants as in other similar research †


Independent Variable

• An *independent variable* is a variable that is manipulated through the design of the experiment
• It is “independent” because it is independent of participant behaviour (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

<table>
<thead>
<tr>
<th>Factor</th>
<th>Test Conditions (Levels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device</td>
<td>mouse, trackball, joystick</td>
</tr>
<tr>
<td>Feedback mode</td>
<td>audio, tactile, none</td>
</tr>
<tr>
<td>Task</td>
<td>pointing, dragging</td>
</tr>
<tr>
<td>Visualization</td>
<td>2D, 3D, animated</td>
</tr>
<tr>
<td>Search interface</td>
<td>Google, custom</td>
</tr>
</tbody>
</table>

Dependent Variable

- A dependent variable is a variable representing the measurements or observations on an independent variable.
- Examples include task completion time, speed, accuracy, error rate, throughput, target re-entries, retries, key actions, 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”)

Three “Other” Variables

- Important but usually given less attention are
  - Control variables
  - Random variables
  - Confounding variables

Control Variable

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

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

Confounding Variable

- Any variable that varies systematically with an independent variable is a confounding variable
- Example 1 – three techniques are compared (A, B, C)
  - All participants are tested on A, followed by B, followed by C
  - Performance might improve due to practice
  - “Practice” is a confounding variable (because it varies systematically with “technique”)
- Example 2 – two search engine interfaces are compared (Google vs. new)
  - All participants have prior experience with Google, but no experience with the new interface
  - “Prior experience” is a confounding variable
Within Subjects, Between Subjects

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

Within vs. Between Subjects

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

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

<table>
<thead>
<tr>
<th>3 X 3 Latin Square</th>
<th>4 x 4 Latin Square</th>
<th>4 x 4 Balanced Latin Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C</td>
<td>A B C D</td>
<td>A B C D</td>
</tr>
<tr>
<td>B C A</td>
<td>B C D A</td>
<td>B D A C</td>
</tr>
<tr>
<td>C A B</td>
<td>C D A B</td>
<td>D C B A</td>
</tr>
<tr>
<td></td>
<td>D A B C</td>
<td>C A D B</td>
</tr>
</tbody>
</table>

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

- Counterbalancing using a Latin Square requires \( m \) participants, where \( m \% n = 0 \) (\( n \) is the number of test conditions)
- Sometimes this is not practical or possible (e.g., number of participants is unknown)
- Alternatively, learning effects may be minimized by
  - Randomizing the order of presentation
  - Using “all possible orders”; e.g., \( n = 3 \rightarrow \) ABC, ACB, BAC, BCA, CAB, CBA

Succinct Statement of Design

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

- Observe and measure
- Research questions
- User studies – group participation
- User studies – terminology
- User studies – step by step summary
- Parts of a research paper

Steps in Empirical Research (1)

Phase I – The Prototype

Think, Analyse, Model, Create, Choose, etc.  →  Build Prototype  →  Test, Measure, Compare  →  Short paper, Poster, Abstract

Iterations are frequent, unstructured, intuitive, informed, ...

Research questions "take shape" (i.e., certain measurable aspects of the interaction suggest "test conditions”, and “tasks” for empirical inquiry.)
Steps in Empirical Research (2)

Phase II – The User Study

- **Build Apparatus** (integrate prototype and test conditions into experimental apparatus & software)
- **Experiment Design** (tweak software, establish experimental variables, procedure, design, run pilot subjects)
- **User Study** (collect data, conduct interviews)
- **Analyse Data** (build models, check for significant differences, etc.)
- **Publish Results**
- **Next iteration**

Themes

- Observe and measure
- Research questions
- User studies – group participation
- User studies – terminology
- User studies – step by step summary
- Parts of a research paper
Research Paper

• The final step
• Research is not finished until the results are published!

Organization of a Research Paper

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

Formatted according to submission requirements of conference or journal (e.g., click here view template for CHI submissions).
Example Publication†

Abstract

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

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

• Tell the reader what you did and how you did it.
• The research must be reproducible!
• Use the following subsections…
Method - Participants

Ten paid volunteer participants (8 male, 2 female) were recruited from the local university campus. Participants ranged in age from 20 years to 40 years (mean = 30.1, sd = 8.5). All were daily users of computers, reporting 3 to 12.5 hours of usage per day (mean = 7.6, sd = 3.5). Self-assessed typing speeds ranged from 35 to 105 words per minute (mean = 62.7, sd = 22.5). Six users described themselves as “regular users of computer games.”

Method - Apparatus

The experiment was conducted in a quiet office using a 400 MHz Pentium-class desktop computer running under Microsoft Windows 98. The system included a 19” colour monitor and a standard mouse and keyboard. The default keyboard mapping for the input keys was Left-arrow = Z, Right-arrow = X, Select = Enter. During entry, the middle finger and index finger on the left hand pressed the Left and Right arrow keys, respectively, while the index finger on the right hand pressed the Select key. These movements were used to control the cursor.
York University – Department of Computer Science and Engineering

Method - Procedure

Procedure
Participants completed a pre-test questionnaire soliciting demographic and computer usage information (revisited above) and a post-test questionnaire on their subjective impressions of the methods (discussed later). Prior to collecting data, the experimenter briefly explained the task and demonstrated the software. The instructions were to enter a series of text phrases "as quickly and accurately as possible" using the specified input technique. Participants were instructed to ignore mistakes and try again.

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

Method - Design

The experiment was a within-subjects design with two conditions: Method 82 vs. Method 66. The order of conditions was counterbalanced. Half the participants entered text first using Method 82, then using Method 66. For the other half, the order was reversed. The software recorded a timestamp and key code for each keystroke, saving these in files for follow-up analyses.

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

• 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

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

Acknowledgment

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

Summary

- Observe and measure
- Research questions
- User studies – group participation
- User studies – terminology
- User studies – step by step summary
- Parts of a research paper

References

4. Card, T. D., Tuomi, C. and Bruce, S. POWER.
Course notes available at...
http://www.yorku.ca/mack/CourseNotes.html

Thank you