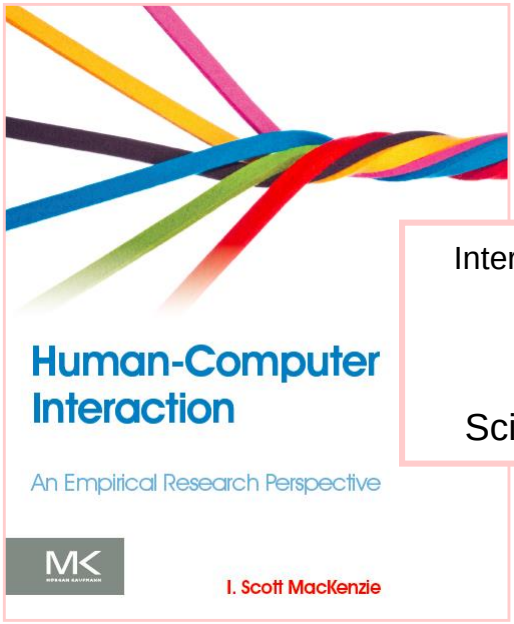


Interactive Systems Design
Prof. V. Fuccella

Chapter 4
Scientific Foundations

Outline

- Research: theory and practice
- Research VS Engineering and Design
- Research Methods
- Scales of Measurement
- Research Questions and Validity of Experimental Procedures

Research – Definition

- Research is…

> Investigation or experimentation aimed at the discovery and interpretation of facts, the revision of accepted theories or laws in light of new facts.

- Example
  - Design and conduct a user study to test whether a new interaction technique improves on an existing interaction technique.

3

Experimentation

- A central activity in HCI research
- An experiment is sometimes called a *user study*
- Formal, standardized methodology preferred
  - Brings consistency to a body of work
  - Facilitates reviews and comparisons between different user studies

4

Research Must Be Published

- Publication is the final step
- Also an essential step
- *Publish or perish!*
  - Edict for researchers in all fields, and particularly in academia
- Until it is published, research cannot achieve its critical goal:
  - Extend, refine, or revise the existing body of knowledge in the field

5

Peer Review

- Research submitted for publication is reviewed by *peers* – other researchers doing similar research
- Only research meeting a high standard of scrutiny is accepted for publication
  - Are the results novel and useful?
  - Does the methodology meet the expected standards for the field?
- Accepted research is published (possibly online) and archived
- The final step is complete

6

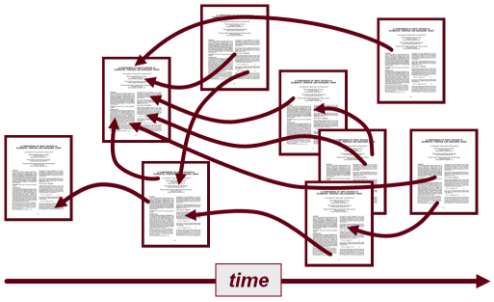
Patents

- Some research develops into bona fide inventions
- A researcher/company may wish to maintain ownership of (profit from) the invention
- Patenting is an option
- The patent application describes
  - Previous related work
  - How the invention addresses a need
  - The best mode of implementation
- If the application is granted, the patent is issued
- Note: A patent is a publication; thus patenting meets the must-publish criterion for research

7

Citations, References, Impact

- Citations, like hyperlinks, connect research to other research
- The number of citations to a research paper is an indication of the paper's impact
- Citations are used to evaluate publication venues (IF, SJR, SNIP)
- Can you spot the high-impact paper below? (arrows are citations)



8

Research Must Be Reproducible

- Research that cannot be replicated is useless
- A high standard of reproducibility is essential
- The research write-up must be sufficiently detailed to allow a skilled researcher to replicate the research if he/she desired
- The easiest way to ensure reproducibility is to follow a standardized methodology
- Many great advances in science pertain to methodology (e.g., Louis Pasteur's detailed disclosure of the methodology used in his research in microbiology)
- The most cited research paper is a "method paper"[1] (see Google Scholar for the latest citation count)

[1] Lowry, O. H., Rosenbrough, N. J., Farr, A. L., & Randall, R. J. (1951). Protein measurement with the Folin phenol reagent. *Journal of Biological Chemistry, 193*, 265-275.

9

Digital Libraries – H-index[1]

- Since the arrival of DL (Google Scholar, Scopus, etc.), citation counts are easy to gather
  - Can be gathered for papers, journals, etc.
  - Can also be gathered for researchers
- H-index is a measure of the impact of a researcher
- Calculation:
  - Rank a researcher's publications by the number of citations
  - the H-index is the point where the rank equals the number of citations;
- A researcher with H-index = *n* has *n* publications each with *n* or more citations

[1] Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*, 16568-16572.

10

H-Index Calculation

Let's suppose a researcher has the publications reported in the table

Title	Citations
Title 1	4
Title 2	0
Title 3	12
Title 4	3
Title 5	5
Title 6	2
Title 7	0

1. We sort them by the number of citations

Title	Citations
Title 3	12
Title 5	5
Title 1	4
Title 4	3
Title 6	2
Title 2	0
Title 7	0

2. We count the rows until position <= # of citations

H-index = 3

11

H-index

- A respectable H-index (although debatable) is "number of years since PhD"
- Exercise: Open Scopus and search total number of papers, total number of citations and H-index of…
  - A professor you appreciate (or you don't)
  - A famous contemporary scientist

[1] Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*, 16568-16572.

12

DL for Computer Science

- Google Scholar – all fields of science
  - https://scholar.google.it/
- ACM Digital Library – computer science
  - http://dl.acm.org/
- IEEE Explorer - engineering
  - http://ieeexplore.ieee.org/Xplore/home.jsp
- Science Direct – several fields of science
  - http://www.sciencedirect.com/

13

Quality of the Venue

- Not all scientific venues have the same prestige.
- It is possible to evaluate journals through:
  - **Impact Factor (IF):** measure reflecting the yearly average number of citations to recent articles published in that journal
  - **Scimago Journal Rank (SJR):** Similar to IF, but citations are weighted on the prestige (SJR) of the citing journal (similar to PageRank).
  - **Source Normalized Impact per Paper (SNIP):** citations received VS citations expected for the journal's subject field.

14

HCI Venues

- Conferences:
  - CHI: The ACM CHI Conference on Human Factors in Computing Systems
  - UIST: Annual ACM Symposium on User Interface Software and Technology
  - IUI: ACM International Conference on Intelligent User Interfaces
- Journals:
  - Human–Computer Interaction
  - TOCHI: ACM Transactions on Computer-Human Interaction
  - International Journal of Human-Computer Studies – Elsevier
  - Interacting with Computers - Oxford Journals
  - IEEE Transactions on Human-Machine Systems

15

Exercise

- Venues are also classified by scientific societies or committees of experts
  - In Italy, classification for computer science conferences were made by GRIN (GRuppo di INformatica)
  - The results can be used for public selections
- Exercise I: find the GRIN classification of the reported HCI conferences
- Exercise II: find the ranking of the reported HCI journals
  - SJR, SNIP
  - Classification of Journals (see e-learning platform)

16

Exercise

- Work in groups
- Use the main scientific search engines to find some (about 15) important papers describing a novel method for «text entry on smart watches» before 2018 (included).
- Evaluate your work:
  - Divide your documents into 2 sets:
    - Relevant: if present in the provided list
    - Not relevant: otherwise
  - Calculate *precision* (the fraction of retrieved documents that are relevant to the topic) and *recall* (fraction of the documents that are relevant to the topic that are successfully retrieved)

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

17

Organizing Literature

Study (1st author)	Number of Keys[a]	Direct/ Indirect	Scanning	Number of Participants	Speed[b] (wpm)	Notes
Bellman [2]	5	Indirect	No	11	11	4 cursors keys + SELECT key. Error rates not reported. No error correction method.
Dunlop [4]	4	Direct	No	12	8.90	4 letter keys + SPACE key. Error rates reported as "very low."
Dunlop [5]	4	Direct	No	20	12	4 letter keys + 1 key for SPACE/NEXT. Error rates not reported. No error correction method.
Tanaka-Ishii [25	3	Direct	No	8	12+	4 letters keys + 4 keys for editing, and selecting. 5 hours training. Error rates not reported. Errors corrected using CLEAR key.
Gong [7]	3	Direct	No	32	8.01	3 letter keys + two additional keys. Error rate = 2.1%. Errors corrected using DELETE key.
MacKenzie [16]	3	Indirect	No	10	9.61	2 cursor keys + SELECT key. Error rate = 2.2%. No error correction method.
Baljko [1]	2	Indirect	Yes	12	3.08	1 SELECT key + BACKSPACE key. 43 virtual keys. RC scanning. Same phrase entered 4 times. Error rate = 18.5%. Scanning interval = 750 ms.
Simpson [24]	1	Indirect	Yes	4	4.48	1 SELECT key. 26 virtual keys. RC scanning. Excluded trials with selection errors or missed selections. No error correction. Scanning interval = 525 ms at end of study.
Koester [10]	1	Indirect	Yes	3	7.2	1 SELECT key. 33 virtual keys. RC scanning with word prediction. Dictionary size not given. Virtual BACKSPACE key. 10 blocks of trials. Error rates not reported. Included trials with selection errors or missed selections. Fastest participant: 8.4 wpm.

[a] For "direct" entry, the value is the number of letter keys. For "indirect" entry, the value is the total number of keys.
[b] The entry speed cited is the highest of the values reported in each source, taken from the last block if multiple blocks.

[1] MacKenzie, I. S. (2009). The one-key challenge: Searching for an efficient one-key text entry method. *Proc ASSETS 2009*, 91-98, New York: ACM.

18

Research vs. Engineering vs. Design

- Researchers often work closely with engineers and designers, but the skills each brings are different
- Engineers and designers are in the business of building things, bringing together the best in *form* (design emphasis) and *function* (engineering emphasis)
- One can image that there is a certain tension, even trade-off, between form and function
- Sometimes, things don't go quite as planned →

20

Form Trumpeting Function

- The photo below shows part of a laptop computer
- The form is elegant – smooth, shiny, metallic
- The touchpad design (or is in engineering?) has a problem
- No tactile sense at the sides of the touchpad

The fix

21

Duct Tape To The Rescue

A true story!

22

Research Milieu

- Engineering and design are about products
- Research is not about products
- Research is narrowly focused
- Research questions are small in scope
- Research is incremental, not monumental
  - Research ideas build on previous research ideas
  - Good ideas are refined, advanced (into new ideas)
  - Bad ideas are discarded, modified
- Products come later, much later

23

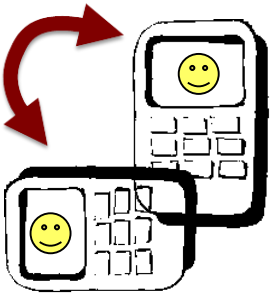
Example: Apple *iPhone* (2007)

iPhone Gestures:

- Tilt
- Flick gesture
- Multitouch

Tilt

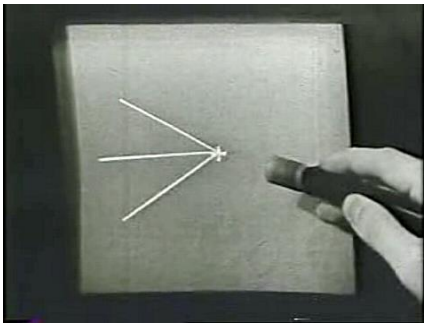
- Research on tilt as an interaction primitive dates at least to 1998[1]



[1] Harrison, B., Fishkin, K. P., Gujar, A., Mochon, C., & Want, R. (1998). Squeeze me, hold me, tilt me! An exploration of manipulative user interfaces. *Proc CHI '98*, 17-24, New York: ACM.

Flick

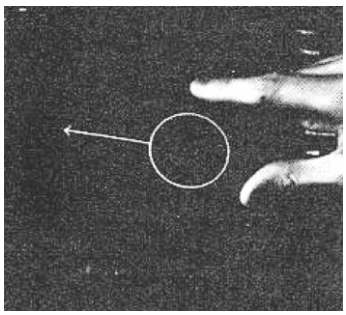
- Research on flick as an interaction primitive dates at least to 1963[1]



[1] Sutherland, I. E. (1963). Sketchpad: A man-machine graphical communication system. *Proceedings of the AFIPS Spring Joint Computer Conference*, 329-346, New York: ACM.

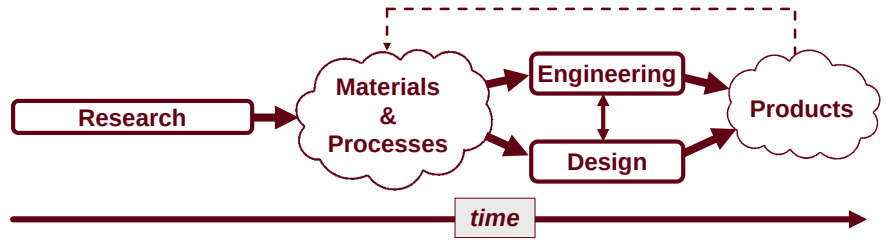
Multitouch

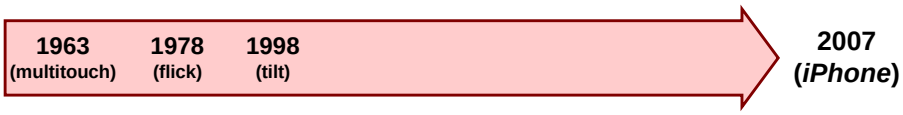
- Research on multitouch as an interaction primitive dates at least to 1978[1]



[1] Herot, C. F., & Weinzapfel, G. (1978). One-point touch input of vector information for computer displays. *Proceedings of SIGGRAPH 1978*, 210-216, New York: ACM.

So…





"Empirical" Research

- Empirical:
  - Originating in or based on observation or experience
  - Relying on experience or observation alone without due regard for system or theory (i.e., don't be blinded by pre-conceptions)
- Example: Nicolas Copernicus (1473-1543)
  - Prevailing system or theory: celestial bodies revolved around the earth
  - Copernicus made astronomical observations that cut against this view
  - Result: heliocentric cosmology (the earth and planets revolve around the sun)

29

Research Methods

- Qualitative VS Quantitative methods

	Qualitative	Quantitative
Objective	Exploratory, understanding things, develop ideas or hypotheses	Used to quantify attitudes, opinions, behaviors
Data collection	Unstructured, data are not numerical	Structured, data are numerical and can be statistically analized.
Sample size	Typically small	Typically large

- Main research methods
  - Observational method (qualitative)
  - Experimental method (quantitative)
  - Correlational method (quantitative)

30

Observational Method

- Focused on qualitative assessments
- Observe phenomena directly in the environment where they occur.
- Example methods:
  - Interviews, field investigations, contextual inquiries, case studies, field studies, focus groups, think aloud protocols, walkthroughs, cultural probes, etc.
- Relevance vs. precision
  - High in relevance (behaviours studied in a natural setting)
  - Low in precision (lacks control available in a laboratory)

31

Experimental Method

- Aka *scientific method*
- Controlled experiments conducted in lab setting
- 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 (changes in the manipulated variable *caused* changes in the response variable)

32

Correlational Method

- Look for relationships between variables
- Observations made, data collected
  - Example: are user's privacy settings while social networking related to their age, gender, level of education, employment status, income, etc.
- Non-experimental
  - Interviews, on-line surveys, questionnaires, etc.
- Balance between relevance and precision (some quantification, observations not in lab)
- Cause-and-effect conclusions not possible

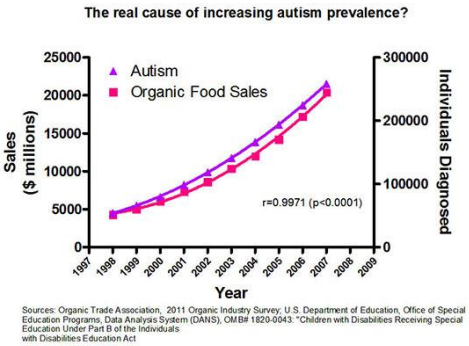
33

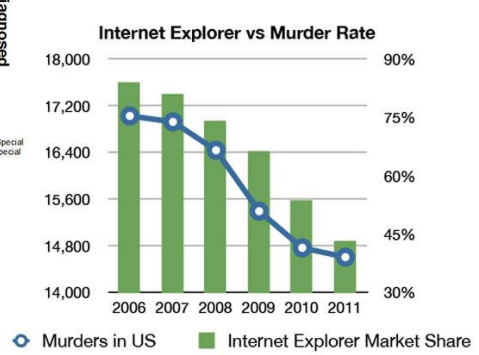
Relationships: Circumstantial & Causal

- As noted above…
  - Correlational methods → circumstantial relationships
  - Experimental methods → causal relationships
- Causal-and-effect conclusions not possible if the independent variable is a *naturally occurring attribute* of participants (e.g., gender, personality type, handedness, first language, political viewpoint)
- These attributes are legitimate independent variables
- But, they cannot be assigned to participants; hence causal relationships not valid
- HCI Example: T9 vs MultiTap (Favourite method)

34

Correlation VS Causation





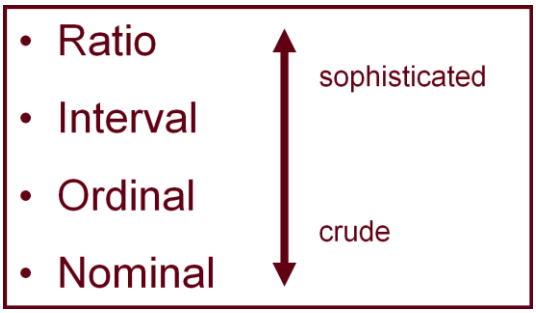
35

Observe and Measure

- Foundation of empirical research
- Observation is the starting point; observations are made…
  - By a human observer
  - By the apparatus
- Manual observation
  - Log sheet, notebooks
  - Screen capture, photographs, videos, etc.
- Measurement
  - With measurement, anecdotes (*April showers bring May flowers*) turn to empirical evidence
  - "*When you cannot measure, your knowledge is of a meager and unsatisfactory kind*" (Kelvin)

36

Scales of Measurement



37

Nominal Data

- Examples of nominal data (aka *categorical data*) : plate numbers, postal codes, job classification, etc.
- Can also be arbitrary codes assigned to attributes; e.g.,
  - 1 = male, 2 = female
  - 1 = mouse, 2 = touchpad, 3 = pointing stick
- The code needn't be a number; i.e.,
  - M = male, F = female
- 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 handers or right handers have more difficulty with…"
  - Note: The count itself is a ratio-scale measurement

Ordinal Data

- Ordinal data associate an order or rank to an attribute
- More sophisticated than nominal data
  - Comparisons of "greater than" or "less than" possible
- The attribute is any characteristic or circumstance of interest; e.g.,
  - Users try three GPS systems for a period of time, then rank them: 1st, 2nd, 3rd choice
- (example on next slide)

Ordinal Data – HCI Example

How many email messages do you receive each day?

1. None (I don't use email)
2. 1-5 per day
3. 6-25 per day
4. 26-100 per day
5. More than 100 per day

Interval Data

- Equal distances between adjacent values
- 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
  - **Problem**: no absolute zero

Interval Data – HCI Example

- 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)
- Likert scale example → (next slide)

Interval Data – HCI Example (2)

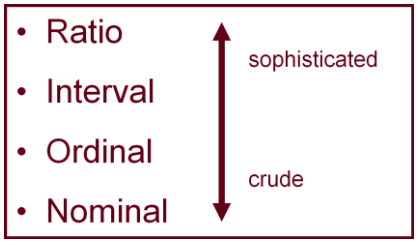
Please indicate your level of agreement with the following statements.

	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 read a text message 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

Ratio Data

- Most sophisticated of the four scales of measurement
- Absolute zero, therefore many calculations possible
- 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
    - Good: Entry rate = 10 / 0.5 = 20 wpm

Scales of Measurement



OK to compute....	Nominal	Ordinal	Interval	Ratio
frequency distribution.	Yes	Yes	Yes	Yes
median and percentiles.	No	Yes	Yes	Yes
add or subtract.	No	No	Yes	Yes
mean, standard deviation, standard error of the mean.	No	No	Yes	Yes
ratio, or coefficient of variation.	No	No	No	Yes

45

Research Questions

- We conduct empirical research to answer (and raise!) questions about UI designs or interaction techniques
- 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?

Testable Research Questions

- Preceding questions, while unquestionably relevant, are not testable
- Try to re-cast as testable questions (even though the new question may appear less important)
- Scenario…
  - You have invented a new text entry technique for touchscreen mobile phones, and you think it's pretty good. In fact, you think it is better than the Qwerty soft keyboard (QSK). You decide to undertake a program of empirical enquiry to evaluate your invention. What are your research questions?

Research Questions (2)

- 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?

Internal Validity

- Definition:
  - The extent to which the effects observed are due to the test conditions
- Statistically, this means…
  - 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 or randomly across the test conditions

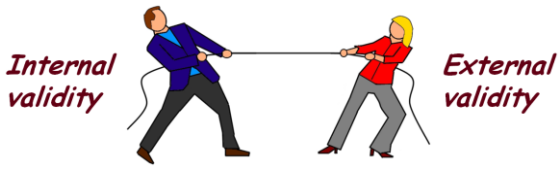
External Validity

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

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 have participants…
  - Enter personalized paragraphs of text (e.g., a paragraph about a favorite movie)
  - Edit and correct mistakes as they normally would
- But… is internal validity compromised?

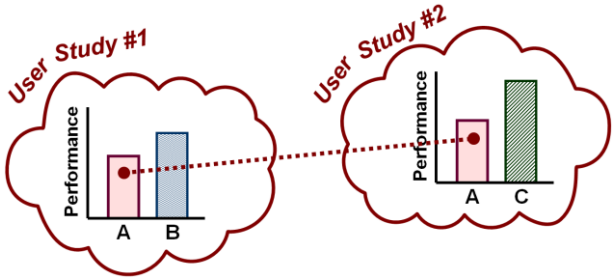
The Tradeoff



- 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, fiddling, or secondary tasks

Comparative Evaluations

- Preferable to do a comparative evaluation rather than one-of
- More insightful results obtained
- Factorial experiments require comparison, because there must be at least one independent variable with at least two levels
- If one condition is a base line; comparisons possible between studies (assuming similar methodology)



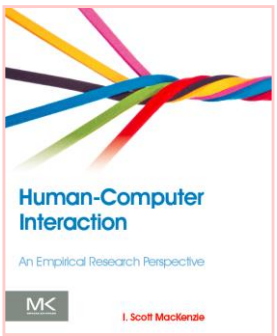
53

Connect to: http://join.quizizz.com

QUESTION TIME

54

Thank You



55