CPT208 Human-Centric Computing

09. Usability Testing & Experimental Design

Dr Yue Li

It’s a great success. Thank you all!

Feedback to your poster presentation

- ➔ Some of you have used related work (academic papers) and your data collection results to clarify your motivation of project. This is a good practice of triangulation.
- ➔ Some groups provided very detailed analysis of existing products.
- ➔ Many groups showed efforts in conducting interviews and questionnaires to understand design requirements.
- ➔ Many groups have specified the user group characteristics (e.g., age, gender, education background, geographical location, technical capacity, etc.). Persona is a good way to help clarify the details.
- ➔ It's great to include some figures and tables to show your results. Make sure the font sizes and colors in your figures are clear to read.

Feedback to your poster presentation

- ➔ Your data (e.g., survey results, comments from target users, etc.) should be used to support your arguments and idea. They should be closely linked with each other. Presenting data with no interpretation or discussion is not meaningful.
- ➔ Use appropriate values to report your data (e.g., means and standard deviations). Also, scale questions often provide more information than yes/no questions.
- ➔ When designing your prototypes, please refer to existing design guidelines (such as Apple's Human Interface Guidelines and Google's material design) to help you get started.
- ➔ Most of you are very well prepared – we are very impressed.

Week 12 – Individual portfolio & video

- The portfolio PDF is a comprehensive document that provides a detailed overview of your project, from conception to evaluation. It should demonstrate your understanding of generative AI, your design and implementation process, and the human-centric impact of your work.

Week 12 – Individual portfolio & video

- Your portfolio should consider including the following content. However, you do not have to follow the exact sequence / structure. For example, you could structure your portfolio based on the key requirements / system features / the design iterations you went through.
  - Title page
  - Background and introduction
  - Design and methodology
  - Prototyping and implementation
  - Evaluation and results
  - Discussion and reflection
  - Conclusion and future work
  - Acknowledgment and references

Week 12 – Individual portfolio & video

- The video is a concise and engaging demonstration of your project, showcasing its key features and functionality. It should be visually appealing and easy to understand, even for viewers unfamiliar with the technical details.

How do we continue?

- We suggest that you
  - sit down together, walk through all your user requirements
  - review your current prototypes, identify areas of improvements
  - agree on your workload distribution
- Examples on how to distribute the workload:
  - On the user requirement level: each of you work on 1-3 functional requirement(s)
  - On the system implementation level: each of you work on improving 1-3 system features
  - On the evaluation level: each of you adopt a different evaluation method to collect and analyze data
- If you do not want to keep working on your group project, you an also start a new project as your individual project

Criteria	Excellent (70+)	Good (60-69)	Satisfactory (40-59)	Needs Improvement (0-39)
Task Fulfilment (70% of 10)	Demonstrates an iterative design process; presents clear prototypes with comparisons; uses appropriate evaluation methods; rich data collection and analysis results; all design decisions are well informed by data and clearly justified; deep discussion on evaluation results; critical reflections on strengths and limitations.	Meets most requirements; shows some iteration and prototype comparisons; uses some evaluation methods; data collection is mostly valid and clearly explained; discussion and reflections are adequate.	Addresses some requirements; limited iteration or prototype comparisons; evaluation methods are unclear; data collection lacks validity or clarity; superficial discussion and reflections.	Fails to meet most requirements; lacks evidence of design process, evaluation, and critical reflection.
Clarity and Organization (30% of 10)	Clear to read, easy to follow, and visually appealing; diagrams, figures, and tables are clear and appropriately used; well-written arguments with no spelling or grammar errors; all sources cited properly.	Mostly clear and organized; minor issues with readability or visuals; some diagrams or tables are unclear; few spelling or grammar errors; most sources cited correctly.	Some clarity and organization; difficult to follow at times; visuals are poorly used or unclear; noticeable spelling or grammar errors; some sources are not cited properly.	Poorly organized; hard to understand; visuals are absent or confusing; frequent spelling or grammar errors; sources are not cited.
Demo Video (100% of 5)	Engaging, clear, and visually appealing; strong prototyping and implementation efforts; demonstrates complete and well functioning system; effectively showcases key features and the design process.	Good video quality; some prototyping and implementation efforts; demonstrates key features with minor issues.	Adequate video; some key features are unclear or poorly demonstrated.	Poor quality; fails to effectively demonstrate key features or the design process.

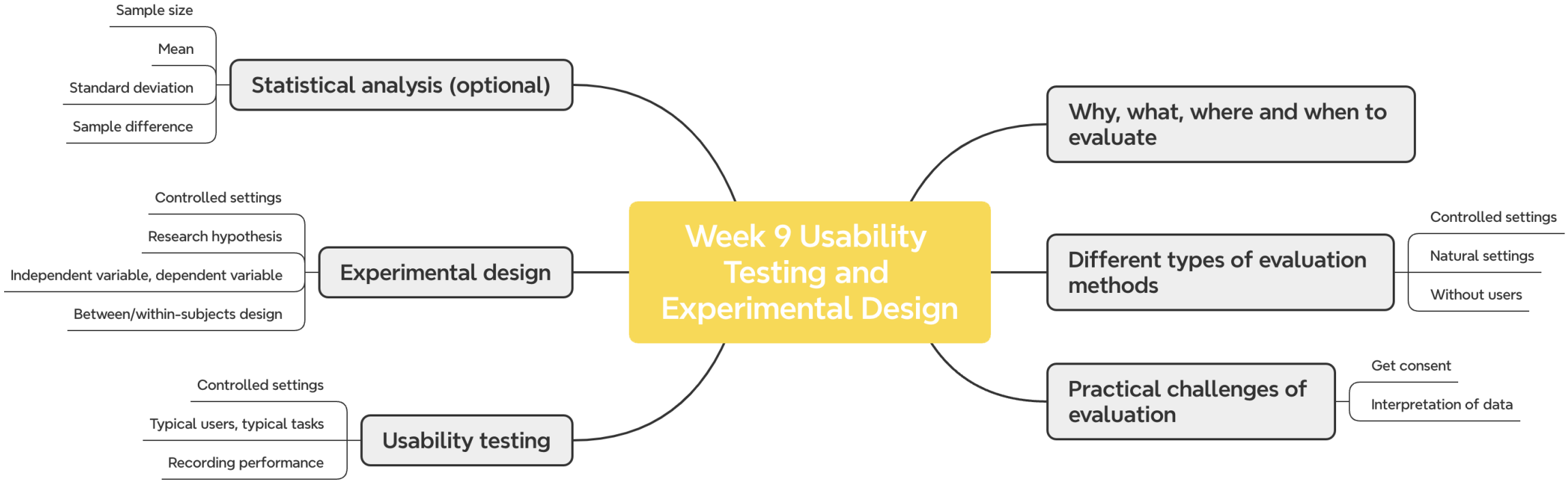
Lecture Syllabus

1. Introduction
2. Discovering requirements
3. Conceptual Prototyping and Practical Guide
4. Design Principles and Design Alternatives
5. Heuristic Evaluation, Questionnaire, and interview
6. Prototyping Fidelity and Dimensions
7. *SAT reading week*
8. *Group Project Demonstration Day*
9. Usability Testing & Experimental Design
10. Interfaces and Research Considerations
11. Field Study and Analytics
12. Flipped Classroom: Selected Coursework Demonstration
13. Revision

Learning Outcomes

1. Understand the why, what, where and when of evaluation
2. Introduce a range of different types of evaluation methods
3. Discuss some of the practical challenges of doing evaluation
4. Explain how to do usability testing
5. Outline the basics of experimental design

This lecture is based on **Chapter 14 and 15** of the ID book.


Week 9 Usability Testing and Experimental Design
Statistical analysis (optional)
Sample size
Mean
Standard deviation
Sample difference
Experimental design
Controlled settings
Research hypothesis
Independent variable, dependent variable
Between/within-subjects design
Usability testing
Controlled settings
Typical users, typical tasks
Recording performance
Why, what, where and when to evaluate
Different types of evaluation methods
Controlled settings
Natural settings
Without users
Practical challenges of evaluation
Get consent
Interpretation of data

Why, what, where and when to evaluate

Chapter 14

Why, what, where, and when to evaluate

Iterative design and evaluation is a continuous process.

- Why: To **check** users' requirements and **confirm** that users can utilize the product and that they like it
- What: A **conceptual model**, early and subsequent **prototypes** of a new system, more **complete prototypes**, and a prototype to compare with **competitors' products**
- Where: In **natural**, in-the-wild, and **laboratory** settings
- When: **Throughout design**; finished products can be evaluated to collect information to inform new products

HCI Research Methods and Measurement

- Early days HCI research measurements were based on human performance (task-based)
  - How fast could someone complete a task?
  - How many tasks were completed successfully?
  - How many errors were made?

HCI Research Methods and Measurement

- Early days HCI research measurements were based on human performance (task-based)
  - Time performance
  - Task correctness / accuracy
  - Error rate
  - Time to learn and retention over time
  - User satisfaction

Micro-HCI and Macro-HCI (Shneiderman, 2011)

- Time performance
- Task correctness / accuracy
- Error rate
- Time to learn
- Retention over time
- User satisfaction
- …

- Motivation
- Collaboration
- Social participation
- Trust
- Empathy
- …

“The old computing is about what computers can do, the new computing is about what people can do.

- Ben Shneiderman, 2002.

Example: Develop a takeaway app

1. You just joined this company
2. You have some initial ideas
3. You have developed some prototypes
4. You have implemented the first version of your design, GanFanBa v1.0
5. You are asked to design an app for the delivery men, SongFanQu
6. You have iterated the design and launched a new version, GanFanBa v2.0
7. ...

Example: Develop a takeaway app

- Why?
  A. To check requirement
  B. To confirm user experience
- What?
  A. A conceptual model
  B. A prototype
  C. A finished product
- Where?
  A. Lab setting
  B. Natural setting
  C. Without (monitoring) user
- When?
  A. Early stage before design alternatives
  B. During design and development
  C. After product launch

Different types of evaluation methods

Chapter 14

Types of evaluation methods

1. Controlled settings that directly involve users
   - → For example, usability and research labs
2. Natural settings involving users
   - → For instance, online communities and products that are used in public places
   - → Often there is little or no control over what users do, especially in in-the-wild settings
3. Any setting that doesn't directly involve users
   - → For example, consultants and researchers critique the prototypes, and may predict and model how successful they will be when used by users

Evaluation methods

Method	Controlled settings	Natural settings	Without users
Observing	√	√	
Asking users	√	√	
Asking experts		√	√
Testing	√		
Modeling			√

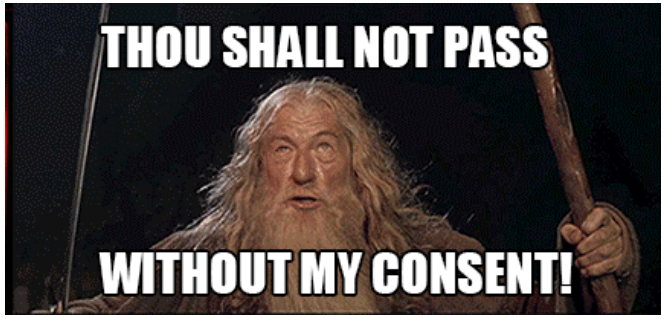
Evaluation methods examples

Method	Controlled settings	Natural settings	Without users
Examples	✓ Usability testing ✓ Experimental design	✓ Field study	✓ Heuristic evaluation ✓ Analytics ✓ A/B testing ✓ Predictive models

Practical challenges of evaluation

Chapter 14

Participants' consent



- Participants need to be told why the evaluation is being done, what they will be asked to do and informed about their rights
- Informed consent forms provide this information and act as a contract between participants and researchers
- The design of the informed consent form, the evaluation process, data analysis, and data storage methods are typically approved by a high authority, such as the Institutional Review Board
  - XJTLU University Ethics Committee (UEC)

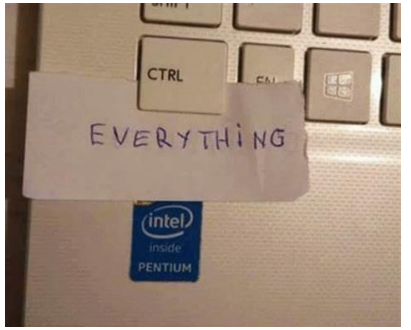
Interpreting data

- Reliability: Does the method produce the same results on separate occasions?
- Validity: Does the method measure what it is intended to measure?
- Ecological validity: Does the environment of the evaluation distort the results?
- Biases: Are there biases that distort the results?
- Scope: How generalizable are the results?

Usability testing

Chapter 15

Usability testing



- Controlled settings
- Users are observed and timed
- Data is recorded on video, and key presses are logged
- The data is used to calculate performance times and to identify and explain errors
- User satisfaction is evaluated using questionnaires and interviews
- Field observations may be used to provide contextual understanding
- Involves recording performance of typical users doing typical tasks

Quantitative performance measures

- Number of users successfully completing the **task**
- Time to complete **task**
- Time to complete task after time away from **task**
- Number and type of errors per **task**
- Number of errors per unit of time
- Number of navigations to online help or manuals
- Number of users making a particular type of error

- Count and calculate data

Usability testing conditions

- Usability lab or other controlled space
- Emphasis on:
  - Selecting representative users
  - Developing representative tasks
- 5-10 users typically selected
- Tasks usually around 30 minutes
- Test conditions are the same for every participant
- Informed consent form explains procedures and deals with ethical issues

How many participants is enough?

- The number is a practical issue
- Depends on:
  - Schedule for testing
  - Availability of participants
  - Cost of running tests
- Typically 5-10 participants
- Some experts argue that testing should continue until no new insights are gained

Example: Computer System Usability Questionnaire (CSUQ)

- Based on Lewis, J. R. (1995) *IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use*. International Journal of Human-Computer Interaction, 7:1, 57-78.
- https://garyperlman.com/quest/quest.cgi

Example: Computer System Usability Questionnaire (CSUQ)

1. Overall, I am satisfied with how easy it is to use this system
2. It was simple to use this system
3. I can effectively complete my work using this system
4. I am able to complete my work quickly using this system
5. I am able to efficiently complete my work using this system
6. I feel comfortable using this system
7. It was easy to learn to use this system
8. I believe I became productive quickly using this system
9. The system gives error messages that clearly tell me how to fix problems
10. Whenever I make a mistake using the system, I recover easily and quickly
11. The information (such as online help, on-screen messages, and other documentation) provided with this system is clear
12. It is easy to find the information I needed
13. The information provided for the system is easy to understand
14. The information is effective in helping me complete the tasks and scenarios
15. The organization of information on the system screens is clear
16. The interface of this system is pleasant
17. I like using the interface of this system
18. This system has all the functions and capabilities I expect it to have
19. Overall, I am satisfied with this system

Example: System Usability Score (SUS)

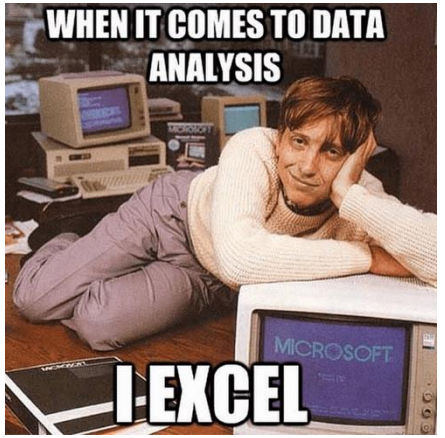
1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Example: System Usability Score (SUS)

To calculate a score between 0 and 100 for the product:

- ➔ Convert SUS responses to numbers, 1 for “Strongly Disagree”, and 5 for “Strongly Agree”.
- ➔ For odd-numbered questions, subtract 1 from the response.
- ➔ For even-numbered questions, subtract the response from 5.
- ➔ Add the scores from each question and multiply the total by 2.5.
- ➔ Remember to present the numbers as a SUS score, not a percentage.

Based on research, a SUS score above a 68 would be considered above average and anything below 68 is below average, however the best way to interpret your results involves “normalizing” the scores to produce a percentile ranking.



More Examples of Established Questionnaires

- Questionnaire for User Interface Satisfaction (QUIS)
- Software Usability Measurement Inventory (SUMI)
- Visual Aesthetics of Website Inventory (VisAwi)
- NASA Task Load Index (NASA TLX)
- ...

Usability Testing and Experiments

- ➔ Usability testing is applied experimentation
- ➔ Developers check that the system is usable by the intended user population by collecting data about participants' performance on prescribed tasks
- ➔ Experiments test hypotheses to discover new knowledge by investigating the relationship between two or more variables

Usability Testing and Experiments

Usability Testing

- Improve products
- Few participants
- Results inform design
- Usually not completely replicable
- Conditions controlled as much as possible
- Procedure planned
- Results reported to developers

Experiments for Research

- Discover knowledge
- Many participants
- Results validated statistically
- Must be replicable
- Strongly controlled conditions
- Experimental design
- Scientific report to scientific community

Experimental design

Chapter 15

Experiments

- Test hypothesis
- Predict the relationship between two or more variables
- Independent variable is manipulated by the researcher
- Dependent variable influenced by the independent variable
- Typical experimental designs have one or two independent variables
- Validated statistically and replicable

Research Hypotheses

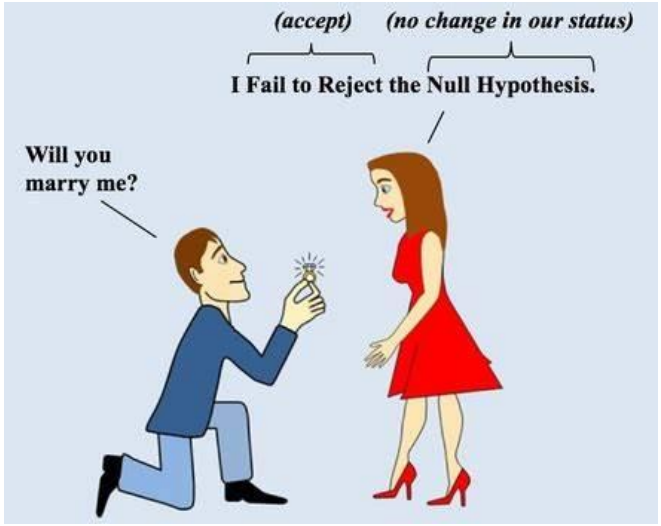
- An experiment normally starts with a research hypothesis
- A hypothesis is *a precise problem statement that can be directly tested through an empirical investigation*

Types of Hypotheses

- Null hypothesis: typically states that there is no difference between experimental treatments
- Alternative hypothesis: a statement that is mutually exclusive with the null hypothesis
- The goal of an experiment is to find statistical evidence to reject the null hypothesis in order to support the alternative hypothesis

Types of Hypotheses

- H0: There is **no difference** in academic performance between students who are in a relationship and those who are not.
- H1: There is **a difference** in academic performance between students who are in a relationship and those who are not.



Experimental designs

- Between subjects design
  - Different participants
  - Single group of participants is allocated randomly to the experimental conditions

- Within subjects design
  - Same participants
  - All participants appear in both conditions

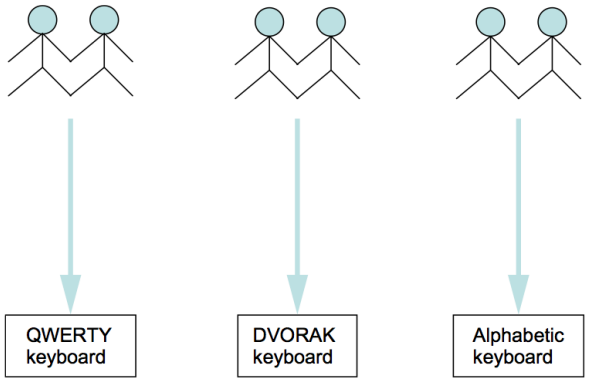


FIG. 3.3
Between-group design.

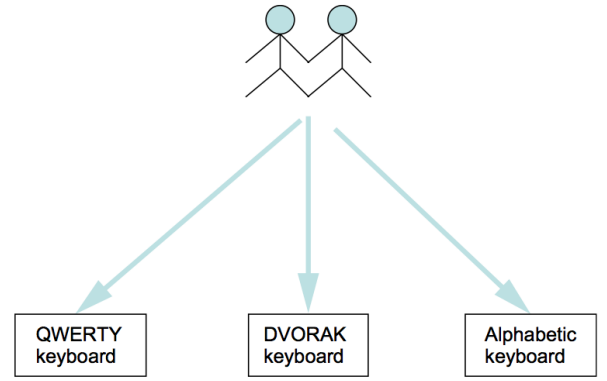


FIG. 3.4
Within-group design.

Experimental designs

Design	Advantages	Disadvantages
Between-subjects	No order effects	Many subjects and individual differences is a problem
Within-subjects	Few individuals, no individual differences	Counter-balancing needed because of ordering effects

Good Practices

- Counterbalancing condition orders through a Latin Square Design.
  - A B C D
  - D A B C
  - C D A B
  - ?
- Provide sufficient training time for users to get acquainted with a system to eliminate the learning effect.
- A single experiment session should be 60-90 minutes or shorter (Nielson, 2005).

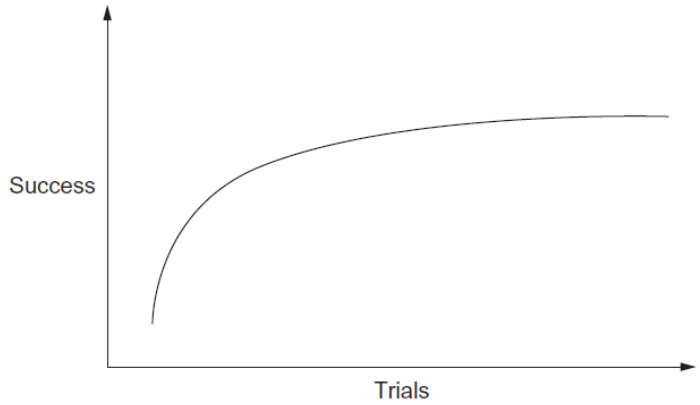


FIGURE 3.5
Typical learning curve.

Experimental design lifecycle

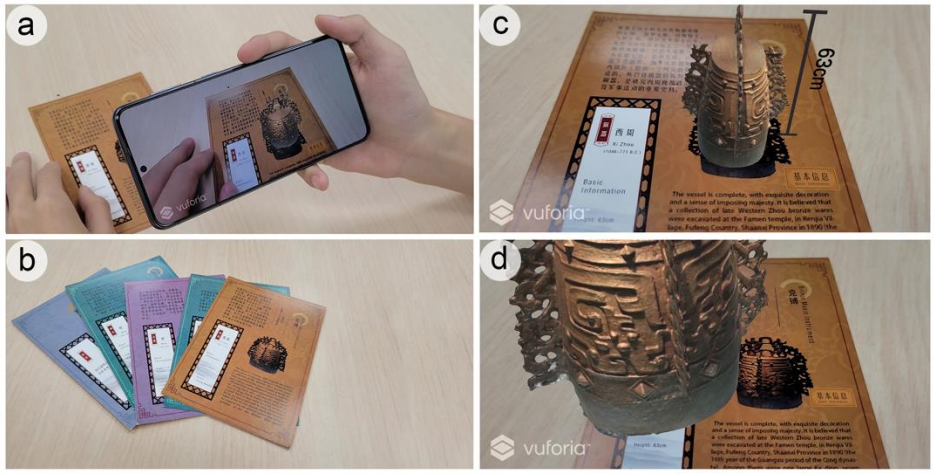
1. Identify a research hypothesis
2. Specify the design of the study
3. Run a pilot study to test the design, the system, and the study instruments
4. Recruit participants
5. Run the actual data collection sessions
6. Analyze the data
7. Report the results

Experiment procedure (Data collection)

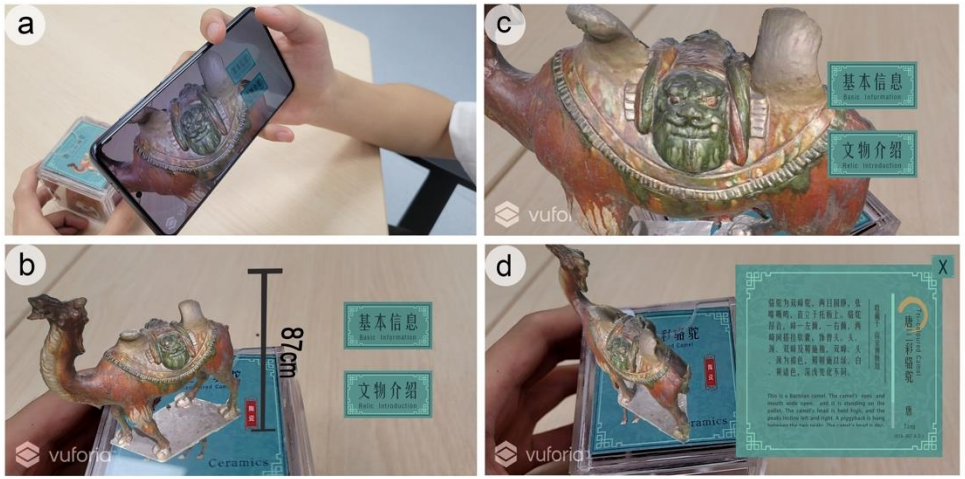
1. Preparation
2. Greet participants
3. Introduce the purpose of the study and the procedures
4. Get consent
5. Assign participants to a specific experiment condition
6. Training task(s)
7. Actual task(s)
8. Participants answer questionnaires (if any)
9. Debriefing session
10. Payment (if any)

Experimental design example: AR prototypes for learning cultural heritage

➔ Goal: to investigate which AR prototype can better motivate users in learning cultural heritage.



Postcard AR



CubeMuseum AR

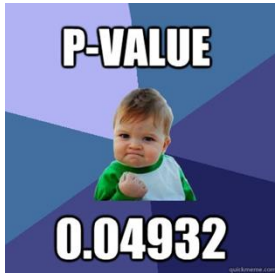
Experimental design example: AR prototypes for learning cultural heritage

- **Independent variable**: two AR prototypes
  - Postcard AR
  - CubeMuseum AR
- **Dependent variable**
  - User motivation
- **Hypothesis**
  - H0: There is no difference in _____ between _____.
  - H1: There is a difference in _____ between _____.

Experimental design example: AR prototypes for learning cultural heritage

Examples of statistical analysis results presentation:

- ➔ There is a statistically significant difference in user motivation between Postcard AR (M=3.04, SD=0.87) and CubeMuseum AR (M=4.32, SD=0.94), t(30)=-4.001, p<0.001.
- ➔ There is no statistically significant difference in user motivation between Postcard AR and CubeMuseum AR, t(30)=-0.740, p=0.465.



Of course, you can use figures and tables to show your data.

Statistical analysis

Additional information

Statistical analysis exercise in Excel

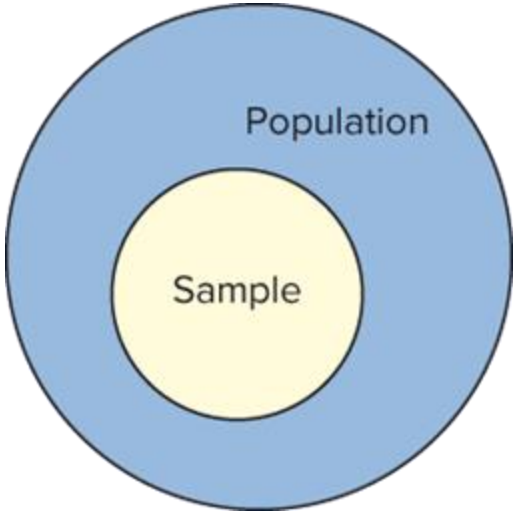
Given the sample data: {5, 5, 4, 3, 4, 1, 5, 4, 4, 4}

- ➔ What is the sample size?
- ➔ What is the mean value?
- ➔ What is the standard deviation?

Given another sample data: {4, 4, 2, 5, 5, 3, 2, 1, 3, 3}

- ➔ How do the two samples differ?

- ➔ See the excel file on LMO for more information.



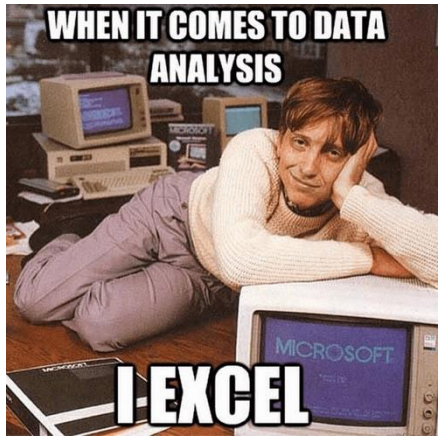
Statistical analysis exercise in Excel

Given the sample data: {5, 5, 4, 3, 4, 1, 5, 4, 4, 4}

- What is the sample size?
  - **`=COUNT(B2:B11)`**
- What is the mean value?
  - **`=AVERAGE(B2:B11)`**
- What is the standard deviation?
  - **`=STDEV.S(B2:B11)`**

Given another sample data: {4, 4, 2, 5, 5, 3, 2, 1, 3, 3}

- How do the two samples differ?
  - **`F-test`** to check variances, then **`t-Test`**.
  - There is no significant difference between the two samples.



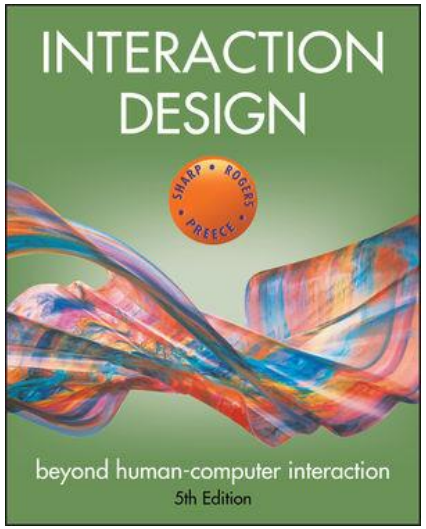
Activity today summary

I hope you have the answers for **your future evaluation studies**:

- Why, what, where, and when to evaluate?
- What type of evaluation methods you've used or plan to use?
  - Controlled? Natural? Without user?
- What exactly are the evaluation methods you've used or plan to use?
  - Usability test? Experimental study? Field study? Heuristic evaluation?
- How did/will you collect, analyse and interpret your data?
  - Questionnaire? Interview? Observation? Statistical analysis? Modelling?

Readings

- Chapter 14: Introducing Evaluation
- Chapter 15: Evaluation Studies: From Controlled to Natural Settings



Any Questions?

Email: yue.li@xjtlu.edu.cn
Office: SD457