HCI: Empirical Research Methods

Learning Objective

- In the previous lectures, we already learned several evaluation methods such as heuristic evaluation, cognitive walkthroughs or cognitive models to evaluate designs at the early phases
- As we have mentioned, interactive system design is not complete unless it is evaluated with end users

Learning Objective

- In this lecture, we shall discuss user evaluation methods
- In particular, we shall discuss the following:
  - The key concerns in user evaluation
  - Data collection procedure
  - Data analysis techniques

Empirical Research

- Empirical research is broadly defined as the "observation-based investigation" seeking to discover and interpret facts, theories, or laws
- Collection and Analysis of end user data for determining usability of an interactive system is an “observation-based investigation”, hence it qualifies as empirical research

Themes of Empirical Research

- Generally speaking, empirical research is based on three themes
  - Answer and raise Questions about a new or existing UI Design or Interaction Method
  - Observe and Measure
  - User Studies

Research Question

- It is very important in an empirical research to formulate “appropriate” research questions
- For e.g., consider some questions about a system
  - Is it viable?
  - Is it as good as or better than current practice?
  - Which of several design alternatives is best?

Research Question

- It is very important in an empirical research to formulate “appropriate” research questions
- For e.g., consider some questions about a system
  - What are its performance limits and capabilities?
  - What are its strengths and weaknesses?
  - How much practice is required to become proficient?

Testable Research Question

- Preceding questions, while unquestionably relevant, are not *testable*
- We have to come-up with testable questions in empirical research

Testable Research Question

- Let’s illustrate the idea with the following example:

  Suppose you have designed a new text entry technique for mobile phones. You think the design is good. In fact, you feel your method is better than the most widely used current technique, multi-tap. You decide to undertake some empirical research to evaluate your invention and to compare it with multi-tap? What are your research questions?

Testable Research Question

- Weak question
  - 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?

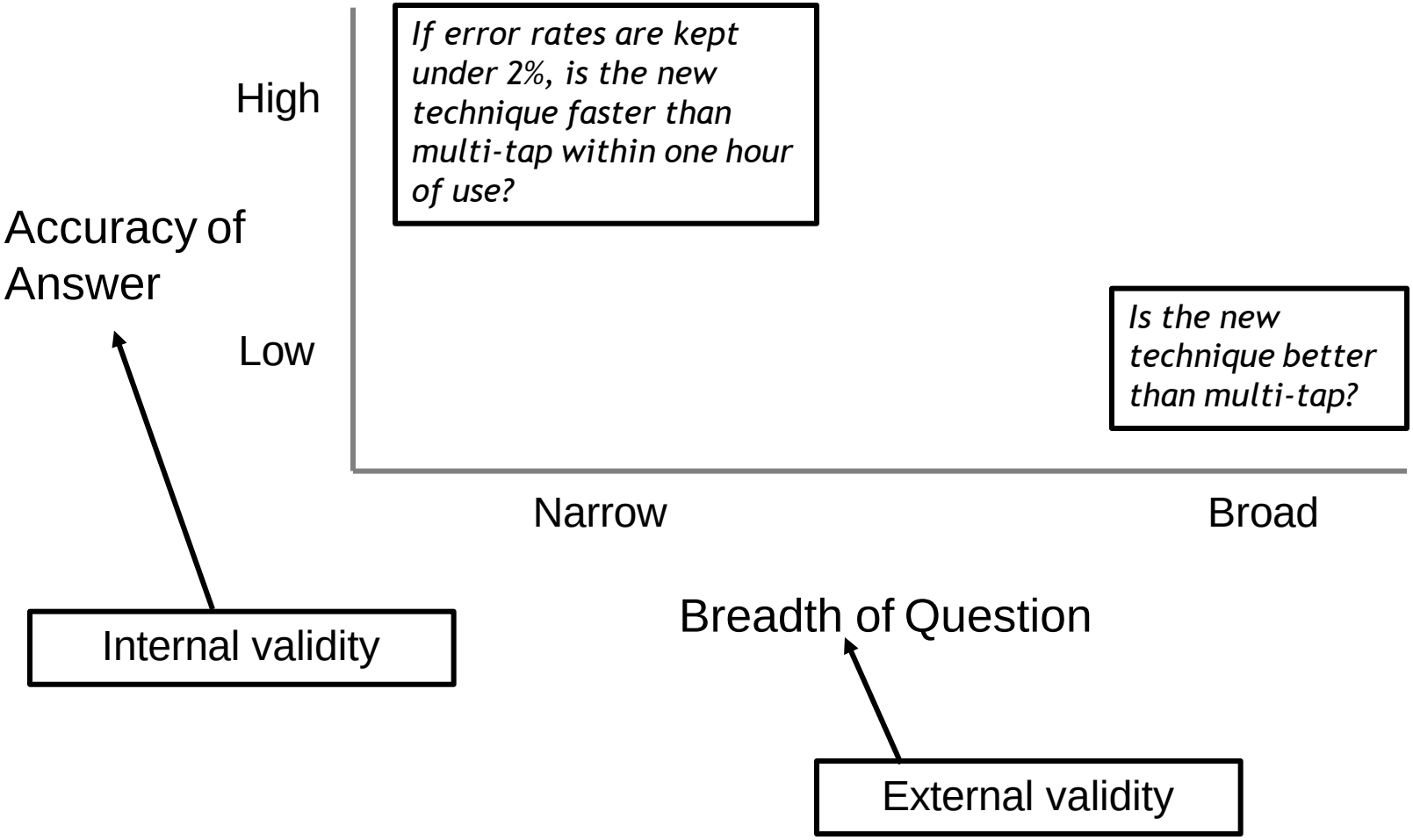
Testable Research Question

- Even better
  - If error rates are kept under 2%, is the new technique faster than multi-tap within one hour of use?
- The questions are testable (we can actually conduct experiments to test the answer to the questions)

Testable Research Question

- We can ask very specific questions (the last one) or relatively broad questions (the first one)
- For very specific questions, the accuracy of answers is high whereas for broader questions, the breadth or generalizability is high

Testable Research Question



Internal and External Validity

- The extent to which the effects observed are due to the test conditions is called internal validity of the research question
- The extent to which results are generalizable to other people and other situations is known as the external validity of the research question

More Examples on Validity

- Suppose 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
  - The test environment should use a projection screen, position participants at a significant distance from screen, have participants stand and include an audience

More Examples on Validity

- Note that creating the test environment mimicking the real usage scenario is not easy
- Instead you can go for controlled experiments where you can ask the user to sit in front of a computer in a laboratory and use the pointing devices to operate an application on the screen
  - The above setting can answer research questions with high internal validity but can not help in determining if the answers are applicable in real world

More Examples on Validity

- Consider another scenario where you wish to compare two text entry techniques for mobile devices
- To improve external validity, the test procedure should require participants to enter representative samples of text (e.g., phrases containing letters, numbers, punctuation, etc.) and correct mistakes
  - This may require compromising on internal validity

Trade-off

- 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

Resolving the Trade-off

- Internal and external validity are increased by posing multiple narrow (testable) questions that cover the range of outcomes influencing the broader (un-testable) questions

Ex: a technique that is ***faster***, is ***more accurate***, takes ***fewer steps***, is ***easy to learn***, and is ***easy to remember***, is generally ***better***

Resolving the Trade-off

- The “good news” is that there is usually a positive correlation between the testable and un-testable questions
  - For example, participants generally find a UI ***better*** if it is ***faster***, ***more accurate***, ***takes fewer steps***, etc.
- The “good news”, in fact, is not so good after all as it raises more confusions

Implication

- The “good news” actually implies we do not need empirical research!!
- We just do a user study and ask participants which technique they preferred
  - Because of the “positive correlation”, we need not take the pain in collecting and analyzing data

Implication

- However, this is not true
- If participants are asked which technique they prefer (a broad question), they'll probably give an answer… even if they really have no particular preference!
  - There are many reasons, such as how recently they were tested on a technique, personal interaction with the experimenter, etc.

Implication

- Therefore, such preferences need not be indicative of the system performance
  - We need to scientifically ascertain the validity of the preferences expressed by the participants, which requires formulation of testable questions

Implication

- Also, with broader questions, we may not get idea about the feasibility or usefulness of the system
  - It is not enough to know if a system is better than another system only but we also need to know "how much better" (for example, it may not be feasible economically to develop a system that is only 5% better than the current system)

Implication

- Seeking feedback from users on broader questions is not very helpful from another perspective
  - It does not help to identify the strengths, weaknesses, limits, capabilities of the design, thereby making it difficult to identify opportunities for improvements

Implication

- Such concerns can be addressed only with the raising of testable research questions
- An important point to note is, in order to test the validity of research questions through observations, we need **measurements**
  - This brings us to the second theme of empirical research, namely to observe and measure

Observe and Measure

- In empirical research, observation is the most fundamental thing to do
- Observational (empirical) data can be gathered in two ways
  - Manual: in this case, a human observer manually records all the relevant observational data
  - Automatic: The observation can also be recorded automatically, through the use of computers, software, sensor, camera and so on

Observe and Measure

- A measurement is, simply put, a recorded observation
- There are broadly four ***Scales of Measurements*** that are used (nominal, ordinal, interval and ratio)
- ***Nominal*:** here, we assign some (arbitrary) codes to attributes of the observational data (for example, male = 1, female = 2 etc.)

Scales of Measurements

- ***Ordinal*:** in this scale of measurement, the observations are ranked (for example, 1st, 2nd, 3rd etc.)
- ***Interval*:** in interval measurement, we consider equally spaced units but no absolute starting point (for example, 20° C, 30° C, 40° C, …)
- ***Ratio*:** this scale of measurement has an absolute starting point (zero) and uses ratios of two quantities (for example, 20 WPM, 30 CPS etc.)

Scales of Measurement

- Nominal
- Ordinal
- Interval
- Ratio

Crude

Sophisticated

Ratio measurements, being the most sophisticated scale of measurement, should be used as much as possible

Ratio Measurements

- As mentioned in the previous slide, ratio scales are the most preferred scale of measurement
  - This is because ratio scales make it convenient to compare or summarize observations
- If you are conducting an empirical research, you should strive to report “counts” as ratios wherever possible

Ratio Measurements

- For e.g., assume you have observed that “ a 10-word phrase was entered by a participant in an empirical study in 30 seconds”. What should you measure?
  - If you measure the “time to enter text” (e.g., t = 30 seconds) as an indicator of system performance, it is a bad measurement
  - However, if you go for a ratio measurement (Entry Rate = 10/0.5 i.e. Entry Rate = 20 wpm), that is much better and gives a general indication of the performance

Ratio Measurements

- Let us consider another example. Suppose in an empirical study, you observed that a participant committed two errors while entering a 50 character phrase
  - If you measure the "number of errors committed" (i.e., n = 2) as an indicator of system performance, it is a bad measurement
  - However, if you go for a ratio measurement (Error Rate = 2/50, i.e. Error Rate = 0.04 = 4%), that is much better and is a more general performance indicator

Summary

- We have discussed two of the three themes of empirical research, namely: (1)Answer and raise Questions about a new or existing UI Design or Interaction Method,

  (2) Observe and Measure

- We shall continue with the third theme of empirical research (i.e. User Studies) in the next lecture