

Human Computer Interaction

UNIT-4

Lecture 1:

Empirical Research Methods in HCI

Mr. Nachiket Sainis / Ms. Reena Saini



**B K Birla Institute of Engineering & Technology,
Pilani**

Introduction to Empirical Research - I

Objective

- In the previous modules, we learned about 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

Objective

- we shall learn about user evaluation methods
- In particular, we shall learn about
 - 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 example, 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?
 - 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
- **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?

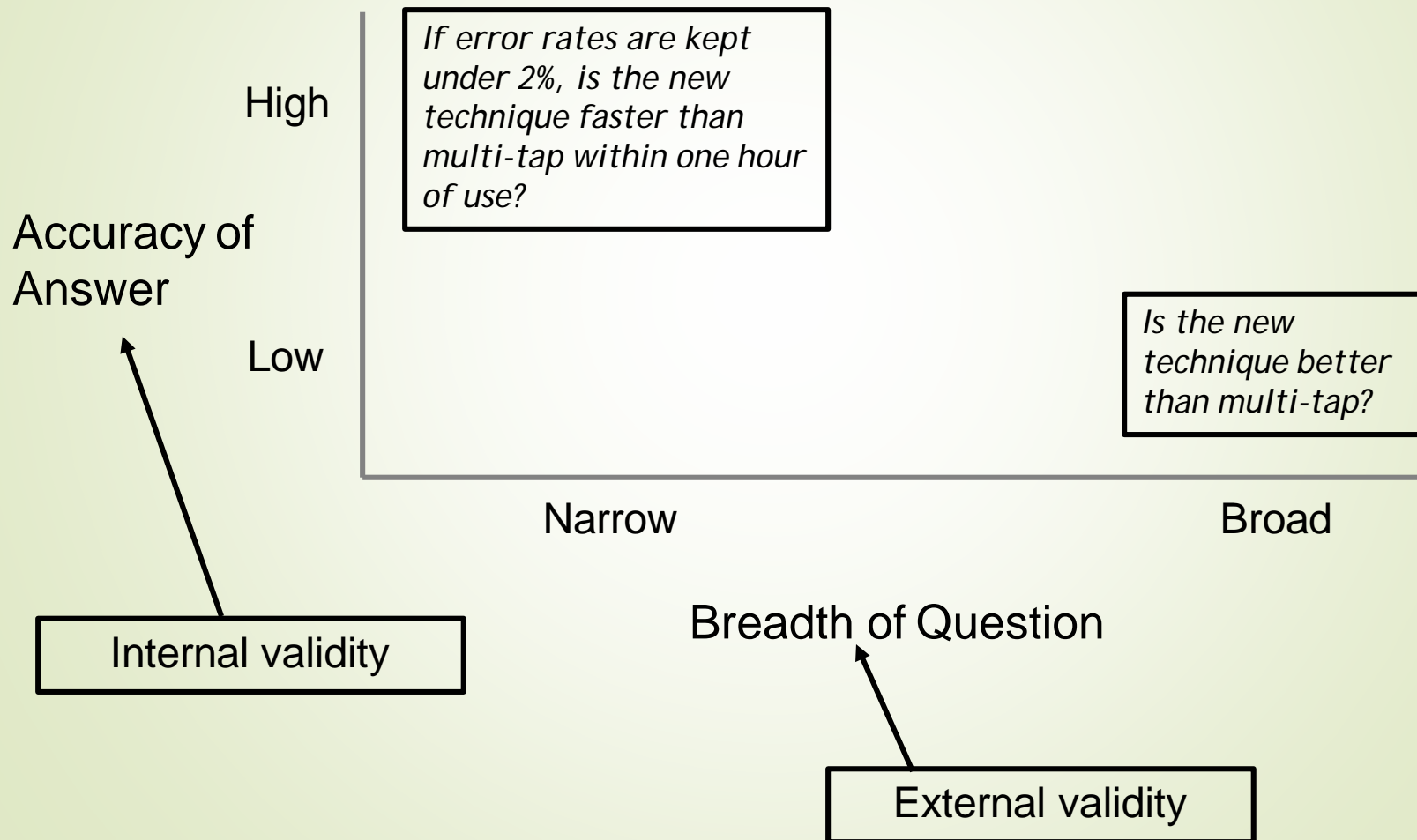
➤ Better still

- 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).
- 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

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

- 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

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
 - 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*
- 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
- 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
- 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

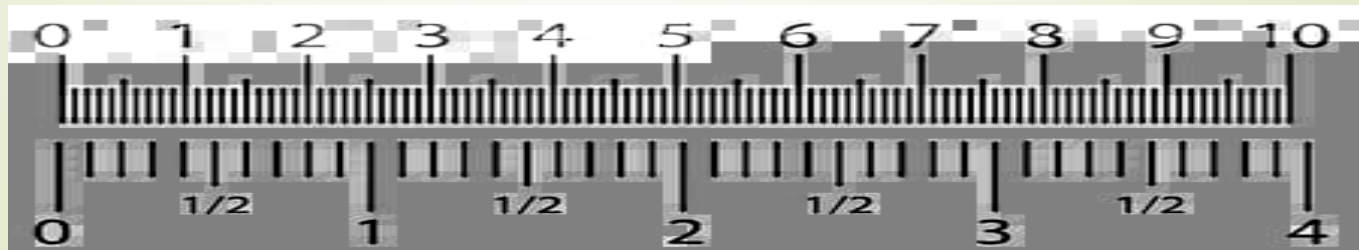
- Seeking feedback from users on broader questions is not very helpful from another perspective
 - It does not help to identify the strengths, weaknesses, limits, capabilities of the design, thereby making it difficult to identify opportunities for improvements
- Such concerns can be addressed only with 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.
- A measurement is, simply put, a recorded observation.
- There are broadly four scales of measurements that are used (nominal, ordinal, interval and ratio).

Scales of Measurements

- Each scale of measurement satisfies one or more of the following properties of measurement.
- **Identity.** Each value on the measurement scale has a unique meaning.
- **Magnitude.** Values on the measurement scale have an ordered relationship to one another. That is, some values are larger and some are smaller.
- **Equal intervals.** Scale units along the scale are equal to one another. This means, for example, that the difference between 1 and 2 would be equal to the difference between 19 and 20.
- **A minimum value of zero.** The scale has a true zero point, below which no values exist.



Nominal Scale

- The nominal scale of measurement only satisfies the identity property of measurement. Values assigned to variables represent a descriptive category, but have no inherent numerical value with respect to magnitude.
- Gender is an example of a variable that is measured on a nominal scale. Individuals may be classified as "male" or "female", but neither value represents more or less "gender" than the other. Religion and political affiliation are other examples of variables that are normally measured on a nominal scale.

What is your gender?	What is your hair color?
<input checked="" type="radio"/> M – Male	<input checked="" type="radio"/> 1 – Brown
<input type="radio"/> F – Female	<input type="radio"/> 2 – Black
	<input type="radio"/> 3 – Blonde
	<input type="radio"/> 4 – Gray
	<input type="radio"/> 5 – Other

Ordinal Scale

- The ordinal scale has the property of both identity and magnitude. Each value on the ordinal scale has a unique meaning, and it has an ordered relationship to every other value on the scale.
- Take a look at the example below. In each case, we know that a #4 is better than a #3 or #2, but we don't know—and cannot quantify—how *much* better it is. For example, is the difference between “OK” and “Unhappy” the same as the difference between “Very Happy” and “Happy?” We can't say.

How do you feel today?

- ☒ 1 – Very Unhappy
- ☐ 2 – Unhappy
- ☐ 3 – OK
- ☐ 4 – Happy
- ☐ 5 – Very Happy

How satisfied are you with our service?

- ☒ 1 – Very Unsatisfied
- ☐ 2 – Somewhat Unsatisfied
- ☐ 3 – Neutral
- ☐ 4 – Somewhat Satisfied
- ☐ 5 – Very Satisfied

Interval Scale

- The interval scale of measurement has the properties of identity, magnitude, and equal intervals.
- A perfect example of an interval scale is the Fahrenheit scale to measure temperature. The scale is made up of equal temperature units, so that the difference between 40 and 50 degrees Fahrenheit is equal to the difference between 50 and 60 degrees Fahrenheit.
- With an interval scale, you know not only whether different values are bigger or smaller, you also know *how much* bigger or smaller they are. For example, suppose it is 60 degrees Fahrenheit on Monday and 70 degrees on Tuesday. You know not only that it was hotter on Tuesday, you also know that it was 10 degrees hotter.



Ratio Scale

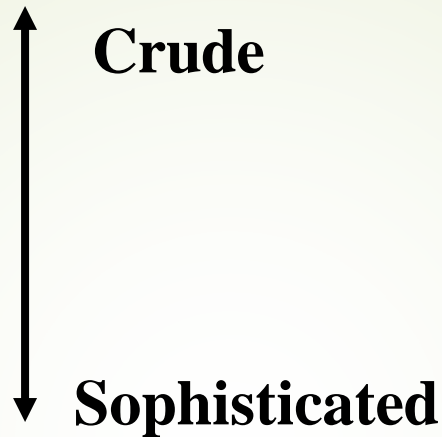
- The ratio scale of measurement satisfies all four of the properties of measurement: identity, magnitude, equal intervals, and a minimum value of zero.
- The weight of an object would be an example of a ratio scale. Each value on the weight scale has a unique meaning, weights can be rank ordered, units along the weight scale are equal to one another, and the scale has a minimum value of zero.
- Weight scales have a minimum value of zero because objects at rest can be weightless, but they cannot have negative weight.

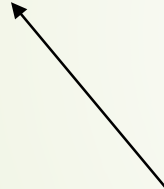


Summary of data types and scale measures

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✓	✓	✓
"Counts," aka "Frequency of Distribution"	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiple and divide values				✓
Has "true zero"				✓

Scales of Measurement

- Nominal
 - Ordinal
 - Interval
 - Ratio
- 
- A vertical double-headed arrow is positioned to the right of the list. The word 'Crude' is at the top of the arrow, and the word 'Sophisticated' is at the bottom. The arrow points upwards from 'Sophisticated' to 'Crude' and downwards from 'Crude' to '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 example, 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” (i.e., $t = 30$ seconds) as an indicator of system performance, it is a bad measurement
 - However, if you go for a ratio measurement (e.g., entry rate = $10/0.5 = 20$ wpm), that is much better and gives a general indication of 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 (e.g., error rate = $2/50 = 0.04 = 4\%$), that is much better and is a more general performance indicator