User-Centric Computing for Human-Computer Interaction NPTEL-MOOCS L27

Dr Samit Bhattacharya

Computer Science and Engineering

IIT Guwahati

Empirical Research Stages

- Broadly, four (or five) stages
  - Identification of research question(s)
  - Determination of variables
  - Design of experiment
  - Analysis of empirical data
  - There is also a fifth stage: building of a model, if that is what we want

Basic Idea

- Once research questions (or hypotheses) are framed, we identify the *variables*
  - To observe and record quantitative data

Basic Idea

- Consider RQ4 we discussed in the last lecture

RQ4: *How the aesthetic score (in a scale of 1-10) depends on the number of objects, object types and the layout of the objects an interface has?*

Basic Idea

- We want to observe aesthetic judgment (of the participants) for RQ4
- So, **aesthetic judgment** is a variable that takes values from the rating scale

Basic Idea

- It can take values between 1 to 10 (integer only) – 10 point rating scale
- Participants give those values (ratings)

Basic Idea

- Interfaces differ w.r.t. number of objects, object types and layout (three variables)

Basic Idea

- First variable (number of objects) can take any integer
- For second variable, we can define types: text only, image only, text and image, text and image with animation and so on - variable can take any of these as value
- Third variable can take as value a layout *specification*

Basic Idea

- Alternatively, for the third variable, we may define a set of different layouts
  - Suppose there are 10 such layouts, numbered 1-10
  - These numbers can also be used as values to the 3rd variable

Basic Idea

- Alternatively, for the third variable, we may define a set of different layouts
  - Suppose there are 10 such layouts, numbered 1-10
  - These numbers can also be used as values to the 3rd variable

Basic Idea

- Once we know variables and their values, we can set up **test conditions** by varying values **systematically**

Basic Idea

- Empirical study done based on these test conditions
- Therefore, identification of variables is necessary
- Along with variables, it is also important to *suitably define* their values

Scales

- In empirical research, we observe and record
- There are broadly TWO ways in which we can record the observations

Scales

- **Manual recording** - human observer records data
- We can also record data **automatically** without human observer – with technology such as computers, sensors, camera, logging software

Scales

- Recording of observed data involve **measurement** of the data

Scales

- Recording of observed data involve **measurement** of the data

Scales

- **Nominal**
  - We **assign some (arbitrary) codes** to attributes of observational data
  - E.g., suppose we wish to record gender data in an empirical study. Instead of recording data in terms of male or female, we may assign the code 1 to male and 2 to female and record the data in terms these two numbers
  - The term **categorical data** is also used

Scales

- **Ordinal**
  - We can **assign some order** on the observed data with this measurement scale
  - Ex- we may observe performance (in terms of playing a game) of three mobile phones and record our observation by ranking phones performance-wise as 1^{st}, 2^{nd} and 3^{rd}

Scales

- **Interval**
  - Recording observations in terms of equally spaced values
  - E.g., If you see a thermometer closely, you might notice some closely spaced marker (lines) indicating specific values (e.g., 98°, 99° and so on). We record temperature based on these markings.

Scales

- **Ratio**
  - Uses ratio of two *quantities*
  - E.g., WPM - ratio of **number of words typed** and **minutes spent on typing**
  - The most sophisticated scale of measurement

Scales

- Ordinal scales are heavily used for *rating*
- A popular rating scale is the Likert scale (named after its inventor who incidentally was a psychologist) [Likert, 1932]

Scales

- **We should strive to use interval or ratio scale as much as possible**
  - These scales support a wide array of analysis techniques as we shall see later

Scales

- Remember that the measurement scales are **important for the observed variables** only

Scales

- In RQ4, one variable is number of objects
- Which scale of measurement we used?
  - We assigned (arbitrarily) an integer (one) to each object and then add them up
  - So, we are using nominal scale
- Same is true with *type of objects* and *layout* variables

Scales

- We formed a research question having three variables that take nominal data
- Is that all right?
- Do we need to change the variables?

Scales

- Answer is NO – in fact, the questions are NOT RELEVANT

Scales

- The three variables are used to generate test conditions - not to observe their values
- We are observing user ratings
- Concern about measurement scales applies to rating variable **only**

Variable Types

- TWO types
  - Dependent variables
  - Independent variables

Independent Variables

- Let us reconsider RQ4
- For experiment, we collect (or design) a set of interfaces
- There are millions of (computer) interfaces in this world - we cannot experiment with all of those
- What we do instead is to select a small sample of those interfaces

Independent Variables

- The sample size is finite - how to choose
  - Through *systematic variation* of variables (number of objects or N, object types or T, object layout or L)

Independent Variables

- Suppose we plan to have two values for N: 4 and 8
  - With these, we wish to see the effect of *less* number of objects on aesthetics (N = 4) and likewise for *more* number of objects in the second case (N = 8)

Independent Variables

- For T, we plan **three values**: text only, image only, text and image together
  - We introduce variations in T for each N
  - In other words, we plan for $2 \times 3 = 6$ interfaces

Independent Variables

- We also consider **two layouts**: *symmetric* and *asymmetric*
  - Applicable for all the six interfaces
  - Therefore, we have altogether 2×3×2 = 12 interfaces for our study

Independent Variables

- We also consider **two layouts**: *symmetric* and *asymmetric*
  - Applicable for all the six interfaces
  - Therefore, we have altogether 2×3×2 = 12 interfaces for our study

Independent Variables

- We can describe these interfaces as,
  - Interface #1 = <N = 4, T = text only, L = symmetric>
  - Interface #2 = <N = 4, T = image only, L = symmetric>
  - Interface #3 = <N = 4, T = image and text, L = symmetric>
- And so on

Independent Variables

- How we arrived at these interfaces - we *assigned* some values to the three variables
  - We *controlled* their values

Independent Variables

- These variables, which are controlled, are known as the *independent* variables
  - Sometimes (in fact, many a times) we use the term *factors* to denote these variables

Independent Variables

- The values assigned to variables are also known as *levels* (of the variable)

Factor	Levels		
Number of objects (N)	4	8	
Object type (T)	Text only	Image only	Text and image
Layout (L)	Symmetric	Asymmetric	

Independent Variables

- Factors combined with their levels give us test conditions for the experiment
  - Ex - “observe aesthetic judgment in a ten point rating scale for the interface having four objects. All these are text objects. The objects are symmetrically organized with respect to the screen coordinate system”
- We can have twelve such test conditions for the factors and levels of Table

Independent Variables

- How do we set the levels for a factor - many ways
  - May come from prior knowledge and experience
  - Theoretical analysis of the problem
  - If all else fails, we might take recourse to a *pilot study* (a small-scale, typically informal and unstructured, study to prepare for the bigger study)

Pilot Study Example

- Ex – deciding on levels for the factor N in RQ4
  - We may *survey* some popular interfaces, record number of objects in those and take average
  - Average may serve as a *threshold* - any value of N below average indicates interface with *less* objects & value above denotes interface with *more* objects
  - Informal survey serves as a pilot study to decide levels for N

Dependent Variables

- In RQ4, there is also a fourth variable: *the user rating* (R)
- R depends on the interface (more specifically, the factors of the interface)
- Therefore, R is called the **dependent** variable

Control & ConfoundingVariables

Control Variables

- In controlled experiment, we control independent variables assuming those to be the only factors that influence observations
  - We may be wrong
- There may be other factors that we are not controlling, which can potentially affect observations

Control Variables

- Ex - we assumed three factors N, T and L in RQ4 the only factors affecting rating
- **What about user gender** - rating of a male user might differ from a female user, for the same interface
- A potential factor but we are not controlling it

Control Variables

- If we know gender is a factor, we may add a fourth factor, gender, to our list of factors

Control Variables

- There is another way of looking at it
  - We may not be interested to know the relationship between the gender and the aesthetic rating
  - In that case, we need not consider it as a factor

Control Variables

- Rather, we will use it as a **control variable** with a fixed value
  - For our experiment, we may use only the male participants
  - Therefore, we set the factor gender to male and use this setting for the whole experiment

Control Variables

- The only problem: any conclusion we draw will be applicable for the male participants only!

Confounding Variables

- Sometimes, we may not even be aware of the existence of factor(s) other than those we already identified (independent and control variables)
- A very common example is the factor called the *practice effect* (or the *learning effect*)

Confounding Variables

- In the experiment for RQ4, suppose a participant sees the interfaces in a particular sequence
- Let there be twelve interfaces
- After seeing and rating first interface, participant gathers some experience in rating and grows some expectations about nature of interfaces
- Experience and expectations get reinforced with each trial (the rating), potentially affecting the way participant rates the interface
- This is the practice effect

Confounding Variables

- Practice effect varies with test conditions systematically although we are not taking into account this variation
- An example of **confounding** variables: we are either not aware of or do not take into consideration their existence in spite of them influencing observations
- We shall learn about a method to deal with the practice effect later

Confounding Variables

- Practice effect varies with test conditions systematically although we are not taking into account this variation
- An example of **confounding** variables: we are either not aware of or do not take into consideration their existence in spite of them influencing observations
- We shall learn about a method to deal with the practice effect later

Why Important

- Consider the problem of predictive aesthetic judgment modeling
  - We identified a testable research question RQ4
  - We also have identified factors and levels – leading to twelve interfaces to study
  - **What should we do now**?

Why Important

- The easiest way out is to rely on your close friends (involves least hassles)
- You ask five of them to participate
- To each, you show the twelve interfaces and ask them to rate
- Ratings are recorded and you get your empirical data

Why Important

- It sounds logical and straightforward
- Unfortunately, there are many issues with the seemingly straightforward approach

Participants - Profile

- Your friends probably belong to same age group and similar background (education, socio-economic conditions and cultural) - a *homogeneous* group
- Older people may have different judgment behavior than teenagers
- It is also likely that judgement is different between a regular computer user and an infrequent user of computers
- A user with a background in arts is likely to judge aesthetics differently than someone without

Participants - Profile

- **What we are talking of is popularly called the *user profile***

Participants - Profile

- User profile may be **factors -** may like to use them as independent variables
- We need to exclusively incorporate them in the final relationship

Participants - Profile

- Downside - you cannot claim the relationship is applicable to *any* user
- **You are modeling behavior of only those users whose profile matches with participants**

Participants - Number

- Other important issue - **number of participants**
  - Is five a good number?
  - Should you go for a bigger number (say ten)?
  - Is that sufficient?

Participants - Number

- There is no definite answer
- According to one estimate, 5 participants good enough for an empirical research
- However, there are alternative views as well

Participants - Number

- In fact, more studies revealed **5 not a good number** and we need to employ more participants
  - Unfortunately, studies could not converge to a single number
- What we get from the literature is a **range** instead, **between 5 and 25**, both inclusive

Participants - Number

- For any pilot study (or small-scale studies), 5 is probably a good number
- However, the conclusions drawn from the data may be treated as **indicative** and more studies may be required

Participants - Number

- To draw *reliable* conclusions from empirical data, we should use **between 12 and 25** participants (more the better)

Note

- Numbers are suggestive only - you can employ more participants
- However, **it is less likely to get different conclusion(s) with more numbers than the ones already got with numbers suggested**

Tasks

- In order to observe, we ask participants to perform *tasks*
  - E.g., in the study on the aesthetic judgment behavior, **task** for the participants is to **rate the interface**
- It is very important to decide appropriate tasks - otherwise, observed data may not be useful

Tasks

- An interface may support large number of tasks
- It is not necessary to ask participants to perform all these tasks - that may be impractical

Tasks

- Instead, participants should be given a set of **representative** tasks
  - Ex – task for VKB design

Assigning Tasks

- Reconsider empirical study for aesthetic judgment based on RQ4
  - 12 test conditions ($2 \times 3 \times 2 = 12$)
  - We should design at least twelve tasks and ask participants to perform these tasks

Assigning Tasks

- Should all participants perform all these tasks?
  - Task is *just* to rate an interface
  - Each participant rates 12 interfaces - not very difficult (hardly takes more than few minutes)
  - Not likely to create any problem if all participants are asked to rate all interfaces

Assigning Tasks

- When each participant in a study performs tasks corresponding to **all** test conditions, we call the study design as "**within-subject**" or "**repeated-measure**"

Assigning Tasks

- We cannot be so always, though
- Let's assume number of test conditions ($5 \times 4 \times 2 = 40$)
- Participants might object (or even refuse) to rate so many interfaces!

Assigning Tasks

- Otherwise, we may ask each participant to rate *only a subset* of the interfaces

Assigning Tasks

- Suppose we have 20 participants
- We divide them into 4 groups (5 in each)
- We ask each group to rate *only* 10 interfaces
- Each participant in a group rates *all* 10 interfaces

Assigning Tasks

- When we distribute the tasks to participants in the way explained above, we call the study as "**between-subject**"

Assigning Tasks

- Within-subject designs are easier to manage - however, practice effect may happen
- It is not easy to decide on the right design (within-subject or between-subject)

Assigning Tasks

- Decision depends on availability of participants (typically, between-subject design requires more participants)
  - Even if participants are available, they might not be able to participate for long (within-subject designs requires longer involvement of the participants)
- Also depends on availability of other resources (study assistants, computers, laboratory space and so on)

Assigning Tasks

- It is necessary to carefully consider these issues and balance the trade-offs

Assigning Tasks

- To avoid **practice effect** in within-subject design, we **counterbalance** sequence in which tasks are given to the participants
  - **Task sequence for each participant is different from the other participants**

Assigning Tasks

- One way - *randomize* the sequence for each participant

Assigning Tasks

- There is another, more systematic, method known as the **Latin Square Method**
- In this method, we organize sequence of tasks given to the participants in the form of a **square matrix**, with the condition that **each task occurs only once in each row and column**

Assigning Tasks

- Ex – let's assume there are 4 participants (numbered P#1 to P#4) for aesthetic judgement behavior study
- Each of them rates 4 interfaces
- Thus, each performs 4 tasks numbered R1 to R4

Assigning Tasks

$$\begin{bmatrix} P\#1 & R1 & R2 & R3 & R4 \\ P\#2 & R2 & R3 & R4 & R1 \\ P\#3 & R3 & R4 & R1 & R2 \\ P\#4 & R4 & R1 & R2 & R3 \end{bmatrix}$$