Lecture 17:
Prototype evaluation (contd.)

Logistics

- Project milestone-2 due soon!
  - Any questions

So far…

- Usability Studies
  - Observe users perform a task and look for deviations from ideal
  - Combine with survey tools.
- Heuristic evaluation
  - Pick a set of heuristics suitable for your goals / tool.
  - Go screen-wise/feature wise
    - For each screen/feature:
      - Evaluate against each heuristic in your set.
  - Go systematically, do not cherrypick!
- Cognitive walkthroughs
  - Walkthrough the interface/task steps as a user would do
  - Use user personas as proxies for the user

Example personas

Name Nisha Aggarwal
Age-26
Working professional

I am a working professional. I hire a porter when I have excess luggage but I am not satisfied with the current booking system because porters do not take genuine price and its very difficult to talk to them. Also I find it uncomofortable to deal with them.

Name: Deepak Kumar
Age-23
College Student

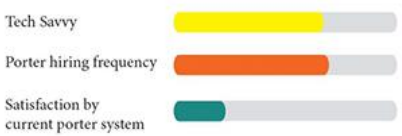
I am a student and I don't hire a porter at Indian Railways most of the time.I usually try to manage my baggage by my self or some times take help from my friends or family if required.I think hiring a porter is too expensive.

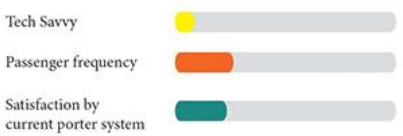
Name: Vinay Sharma
Age- 42
Govt. employee

I travel very frequently. I hire a porter when I have excess luggage but I am not satisfied with the current booking system because porters do not take genuine price and it's a complete waste of time to convince them.

Name Mansha Yadav
Age-32,Porter
High school pass

I am an Indian Railways porter by profession. I am working since last 4 years. Convincing a passenger is difficult for me because they think I demand too much. I spend almost 10-14 hrs daily only for 400-500 rupees. Government price is less so its very hard to serve at government rates.





How are personas created?

- From data – and data only!
- Sift through qualitative data, identify themes
  - Behaviours, attitudes, needs, goals, skills
  - Find categories under each
  - Create personas that capture uniqueness, extremities

CW

- Will <Persona> have formed this goal?
- Will <Persona> have formed this sub-goal?
- Will <Persona> know what to do next / at this stage?
- Will <Persona> do this action?
- Will <Persona> know they did the right thing?

Experiment design: Identifying variables

1. Independent variables (IV) : What researchers change (e.g., interface with vs. without search results) and is independent of other experiment variables

2. Dependent variables (DV) : Depend on independent variables (e.g., search times, no. of search queries, etc.; they might change based on the independent variable (yes/no search history))

   Independent variable → Dependent variable

3. Confound variables (CV): Anything else (other than independent) that might alter the dependent variable in the experiment; ideally, we should not allow confounds to alter results. Atleast, we must reduce their effect on results.

Ensure variables are concretely measurable (e.g., time taken for task completion, and not “productivity”!)

Some examples

- Gas laws in Physics:
  - At Constant Pressure, Volume proportional to Temperature ($V \propto T$)
  - In experiments, we change temperature, and measure volume each time
  - Temperature → independent (experimenter changes)
  - Volume → dependent (on change in temperature)
  - Volume can also depend on pressure, but we don't want it to mess up readings (we are only interested in volume and temperature) → Confound
  - We therefore keep pressure constant (so there is no extra changing effect of pressure across readings)

Some examples

- Using slides in class reduces attention and lowers grades
- Experiment: Half lectures with slides and half without; hand out survey after each lecture on how interesting, did you make notes, were you surprised, did you fall asleep, quiz questions with scores, etc.
  - IV = with/without slides (0 or 1 categorical variable)
  - DV = survey results (quiz scores, interest scores, etc.)
  - Confound? (Due to social desirability bias, power relation between teacher and student, each lecture has different content, different students show up to class each time, etc.)

Experiment design: Study location

- In-vitro=in the lab
  - In the lab; unrealistic, but offers better control (e.g., no distractions)
- In- vivo = in the field
  - Realistic conditions, so higher external validity
  - But, can introduce confounds (e.g., interruptions mess up time measurements, as well as focus and attention)

Experiment design: Participants

- Recruit from user population
- Hard to decide whether to recruit diverse / narrow
- Diverse means controlling for expertise, backgrounds, etc = they might introduce confounds, so we need to ensure the same kinds of people use both systems/prototypes

- How many?
  - Atleast 30 comparisons between the two systems being compared!
  - 30, because many statistical tests operate by comparing distributions of data, and distributions plot smooth at about that size

Study design: Task assignment

- Within-subject
  - Take 2 comparable tasks; Get N participants
  - Each participant does two tasks: one with System1 and one with System2
  - Balance: N/2 participants do Task 1 first, and N/2 do task 2 first.
    N/2 participants do Task1 with System1, and N/2 do Task1 with system2.
  - Compare the difference in DV for both System1 and System2.
  - Question: Why do we need this balancing?
- Between-subject
  - Take one task(s); Get N participants
  - All N participants do same task(s), N/2 with System1 & N/2 with System2
  - Compare average/median/SD between the two groups
  - Confound: Different participants have different skills, motivations, backgrounds, etc.
  - Question: how to deal with confound?

Conclusions from experiment design

- Measurements subject to rigorous statistical tests
- Tests aimed at rejecting a true/false hypothesis
  - There is no difference in mean search times between prototype A and B
  - Tests provide a “p-value” which is basically probability whether any difference observed is “by chance” or is real difference (lower p = lower chance of “by chance”, and higher significant differences). Typically, $p<0.05$.
- They work by comparing distributions of data (e.g., frequency distribution of task time in A vs. B).
- Different tests also make some assumptions about data (sample size, normality of distribution, equal variances, etc.)

- Don’t fret much about specific tests, but this is the general idea!