Lecture 17:
Prototype evaluation (contd.)

Logistics

- Milestone -2 is up
- Hope you are making progress on projects!

Prototype Evaluation

- **With Users**
  - Usability studies
  - Controlled Experiments
- **Without Users**
  - Heuristic Evaluation
  - Cognitive Walkthrough
  - Expert consultation

- **Single System**
  - Usability studies
  - Heuristic Evaluation
  - Cognitive Walkthrough
  - Expert Consultation
- **Comparisons of >1**
  - Controlled experiments

Last class

- Usability Studies (Even N=5 is super helpful!)
- Have users perform a task
- Observe as they perform the task
- Look for confusion, unexpected events
  - Ideal sequence: Insert → SmartArt → Hierarchy → Tree
  - User sequence: Start reading each option item by item, or read smart art and ignored clicking: Sign that something is unusable.
- Ask why the user did what they did (e.g., what did you think the button did → and then use that to debug your design).

You could also use standard questionnaires:

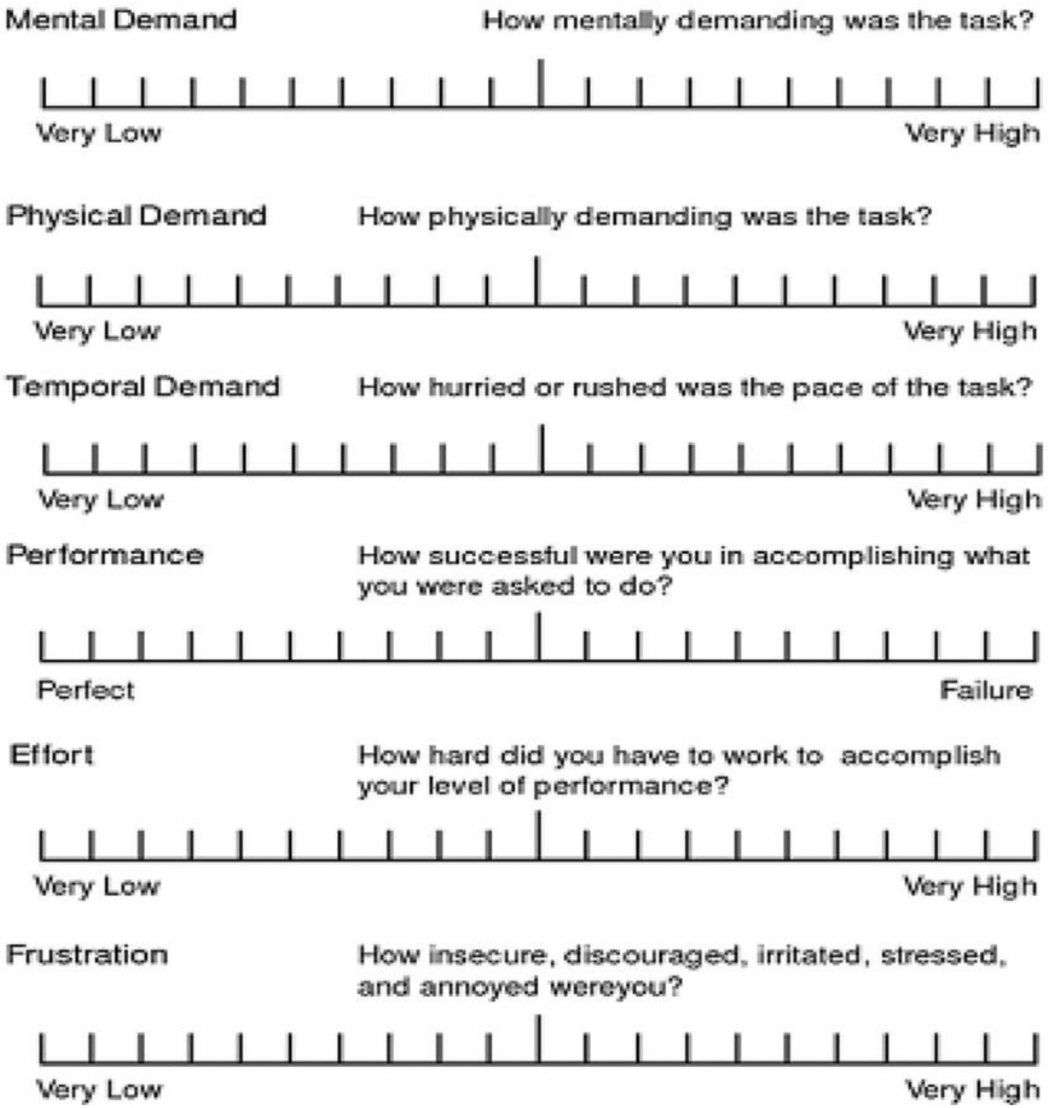
System Usability Scale (SUS)

UXtweak

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

- What’s good here?
- What’s bad here?
- When is a good time to use this?

NASA’s task load index (TLX)


Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

- What’s good here?
- What’s bad here?
- When is a good time to use this?

Heuristic evaluation

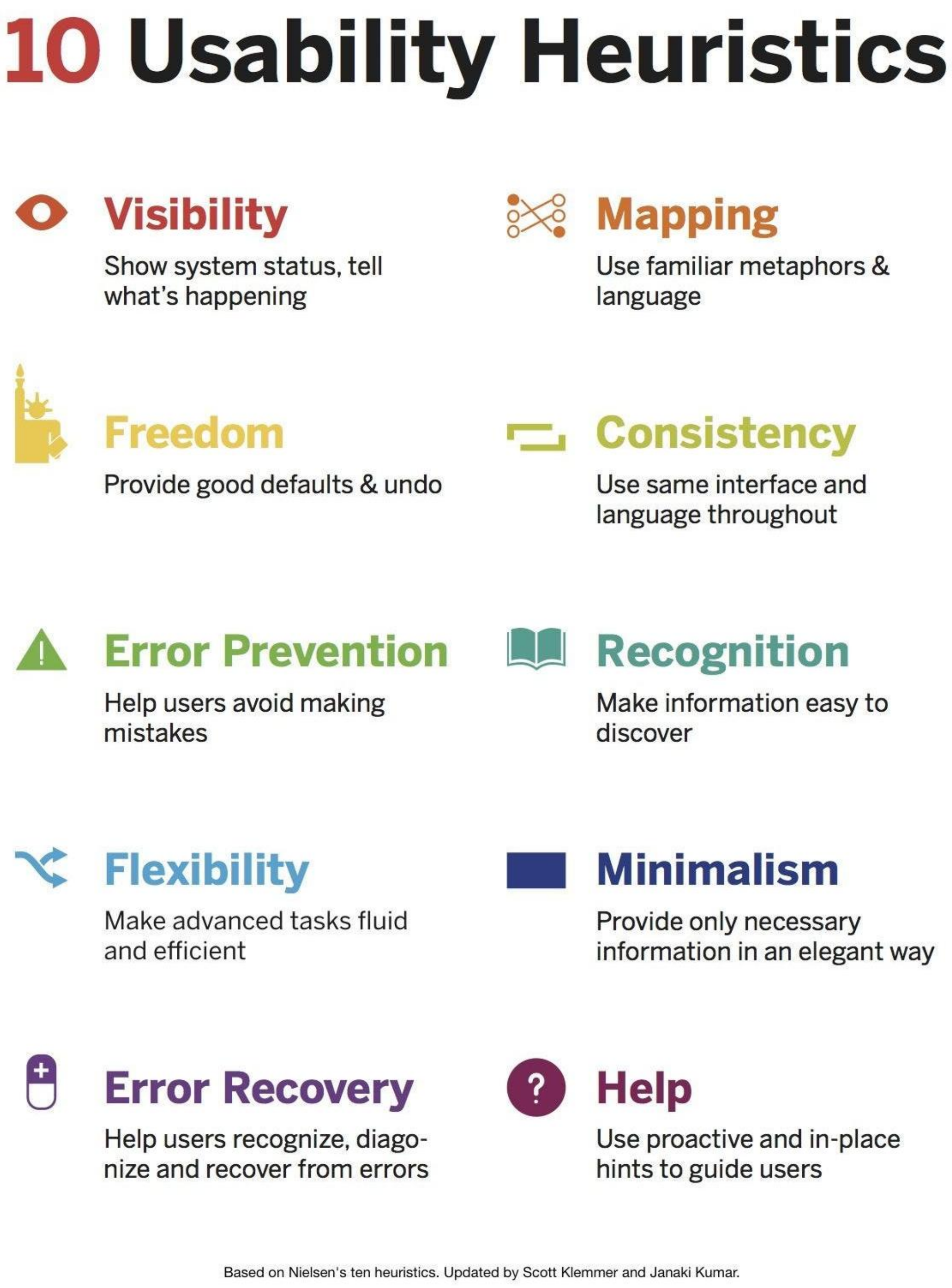
- Heuristics = rules of thumb (for how to build interfaces)
- We evaluate interfaces against heuristics and look for violations
- Result is a list of violations to be fixed
- Who does it?
  - Someone that can interpret UI/UX heuristics, and catch violations (so a UI/UX expert/professional)
  - How many? Ideally, 5 or more.

How to do it?

- Give a set of heuristics
- Give your interface
- These people will find violations of the heuristics in your design.
  - Systematic, and documented one at a time.
  - So also give them a template to fill in.
- What is good?
  - Easy when users are unavailable. Anyone can do it, once they learn the heuristics, helpful as first cut.
- What is bad?
  - Bound to heuristics. Experts are still not users.

Example: IITK Webmail

- Pick a heuristic:
  - E.g., Visibility of System Status
  - List out issues (violations)
  - List out fixes
  - (This is why an expert is needed!)
- Recommended template:
  - https://media.nngroup.com/media/art ation_Workbook_-_Nielsen_Norman_C



Your turn… IITK Web mail

1

Visibility of System Status

The design should always keep users informed about what is going on, through appropriate feedback within a reasonable amount of time.

- Does the design clearly communicate its state?
- Is feedback presented quickly after user actions?

Issues	Recommendations

Your turn… IITK Web mail

4

Consistency and Standards

Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform and industry conventions.

- Does the design follow industry conventions?
- Are visual treatments used consistently throughout the design?

Issues	Recommendations

Example heuristics

- https://www.nngroup.com/articles/ten-usability-heuristics/
- Ben Shneiderman's eight golden rules of interface design
- Microsoft's guidelines for human-AI interaction design
- UI tenets and traps
- Rules from Steve Krug's "Don't make me think" for web usability
- There's a lot more, go look for them!

Cognitive walkthroughs

- Done within the team, when access to users is hard
- Even otherwise, do this as a first cut evaluation
- Pick a prototype and task; list the task steps
- Create a user "persona" (and list down their key characteristics)
  - Often, you need more than persona, then use ones at extremes
- For each step along the task, answer the following 4 questions
  - Will "User" want to do this? [Ideally, use persona name instead of "User"]
  - Assume "User" wants to do this, will s/he know what to do?
  - Assume "User" knows what to do, will s/he actually do it?
  - Assume "User" did it, will s/he know they did the right thing?
- Write down yes/no/maybe, along with reasons. Every no/maybe, is a usability issue to be fixed. The reasons often provide hints for what the fix is.
- Seems tedious, but can be done in an afternoon for atleast most common/least common paths as needed.

A note on personas

- Good personas are data driven, and come from user research (in the empathize phase)
- Example for how to do it, if you care:
  - https://uxpressia.com/blog/how-to-create-persona-guide-examples

- There are a lot of personas out there for use (people with disability, specific problem solving aspects, etc.)

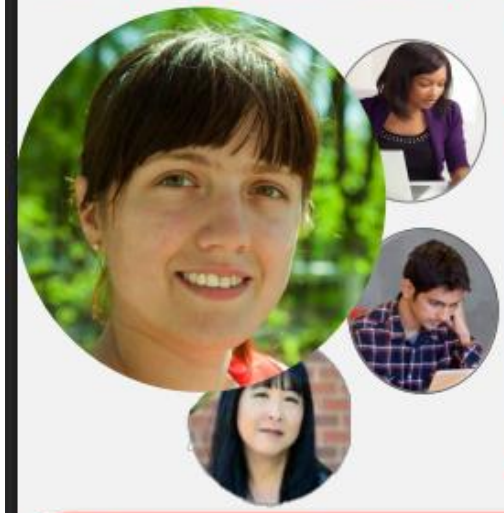
- The definitive guide on the topic is:
  - “The persona lifecycle” by Tamara Adlin and John Pruitt.

Example

Abi (Abigail/Abishek)[1]

Age:
Employment/Position:
Location:
Pronouns:

Abi likes scanning all their emails first to get an overall picture before answering any of them.



Background and Skills

The technologies at Abi's new position are new to them. Abi likes Math and working with logic. They considers themselves a numbers person.

Motivations and Attitudes

- ***Motivations***: Abi uses technologies to accomplish their tasks. They learn new technologies if and when they need to, but prefers to use methods they are already familiar and comfortable with, to keep their focus on the tasks they care about.
- ***Computer Self-Efficacy***: Abi has lower self confidence than their peers about doing unfamiliar computing tasks. If problems arise with their technology, they often blame themselves for these problems. This affects whether and how they will persevere with a task if technology problems have arisen.
- ***Attitude toward Risk***: Abi's life is a little complicated and they rarely have spare time. So they are risk averse about using unfamiliar technologies that might need them to spend extra time on, even if the new features might be relevant. They instead performs tasks using familiar features, because they're more predictable about what they will get from them and how much time they will take.

How Abi Works with Information and Learns:

- ***Information Processing Style:*** Abi tends towards a comprehensive information processing style when they need to gather more information. So, instead of acting upon the first option that seems promising, they gather information comprehensively to try to form a complete understanding of the problem before trying to solve it. Thus, their style is "burst-y"; first they read a lot, then they act on it in a batch of activity.
- ***Learning: by Process vs. by Tinkering:*** When learning new technology, Abi leans toward process-oriented learning, e.g., tutorials, step-by-step processes, wizards, online how-to videos, etc. They don't particularly like learning by tinkering with software (i.e., just trying out new features or commands to see what they do), but when they do tinker, it has positive effects on their understanding of the software.

1Abi represents users with motivations/attitudes and information/learning styles similar to hers. For gender distribution data on users similar to and different from Abi, see http://gendermag.org

*Scenario name:*____

(e.g., Boss just called Abby and told her to remove Kelly's access to the system)

<table>
<tr><td colspan="3">• Subgoal #:
• Subgoal name: (eg, make Kelly not be able to log on)

Analysis questions:
• Will <persona> have formed this sub-goal as a step to their overall goal? YES NO MAYBE (Choose one)
• Why? Especially consider <persona>'s Motivations & Strategies</td></tr>
<tr><td>• Action #: __
• Name:
(eg, click "new" button):</td><td>• Will <persona> know what to do at this step?
YES NO MAYBE
• Why? Especially consider <persona>'s Knowledge/Skills, Motivations/Strategies, Self-Efficacy and Tinkering)</td><td>• If <persona> does the right thing, will s/he know that s/he did the right thing and is making progress toward their goal?
YES NO MAYBE
• Why? Especially consider <persona>'s Self-Efficacy and Attitude Toward Risk</td></tr>
<tr><td>• Action #: __
• Name:
(eg, click "new" button):</td><td>• Will <persona> know what to do at this step?
YES NO MAYBE
• Why? Especially consider <persona>'s Knowledge/Skills, Motivations/Strategies, Self-Efficacy and Tinkering)</td><td>• If <persona> does the right thing, will s/he know that s/he did the right thing and is making progress toward their goal?
YES NO MAYBE
• Why? Especially consider <persona>'s Self-Efficacy and Attitude Toward Risk</td></tr>
</table>

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 <u>comparable</u> 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!