CMPT 281

Evaluations with Users

Learning Objectives

- When & Why to Involve Users in the Design Process
- Get exposure to typical methods:
  - Observation
  - Interviews
  - Questionnaires
  - Participatory design
  - A-B testing
  - Tree testing
  - First click testing

User Centered Design

System-Centered Design

- What can be built easily on this platform?
- What can I create from the available tools?
- What do I as a designer find interesting to work on?

User-Centered Design

- Coined by Don Norman
- Focuses on the user at every stage of design
- Aims to adapt systems to work for users rather than force users to adapt their behavior to the system

Intro to Evaluation

Evaluation: Key Points

- Evaluation should occur throughout the design process
- There are many different evaluation methods
  - Can be used at any time during process
- Evaluation is often equated with usability testing
  - Can only be used the end of the design process
  - But, only one use/type of evaluation!

Evaluation at Various Stages in the Usability Engineering Life Cycle

- Pre-design – needs assessment
  - Viability proof for investment in new expensive system
  - (Is there a real need, and what should the approach be?)
- Initial design stages – user input
  - Develop and evaluate initial design ideas with the user

Evaluation at Various Stages in the Usability Engineering Life Cycle

- Iterative design – user feedback
  - Does system behaviour match the user's task requirements?
  - Are there specific problems with the design?
  - Verify that interface meets expected performance criteria
    - Ease of learning, usability, user's attitude, performance criteria:
    - "A first time user will take 1-3 minutes to learn how to withdraw $50 from the automatic teller"
- Acceptance testing – after all that, does user use it

Why Study Users Before You Design?

- Because you need to know things like:
  - How users do tasks now
  - Current problems: inefficiencies, frustrations, lack of or confusing functionality
  - Current dependencies: what parts of the current system are valid, and need to be retained?
  - If you have an approach for a new design, is it generally likely to solve existing problems?
- To answer this, you have to understand existing problems

Empirical Methods of Directly Studying Users

- Sample surveys / questionnaires / interviews
  - Ask people to report on themselves
- Field studies / observation methods
  - Observe naturally occurring systems
- Experiments and quasi-experiments
  - Observe/measure under controlled conditions

I'm an aspiring web designer/developer – why do I care?

I'm an aspiring web designer/developer – why do I care?

On a smaller scale: useful for testing your own wireframes, prototypes, and final websites – can use many of these methods (or components of them) with clients, friends, and potential customers

- Saves you time in the long run
- Attention to detail = successful websites
- Learn what is generally successful, and what isn't

I'm an aspiring web designer/developer – why do I care?

On a larger scale, ALL major websites make use of some combination of these methods.

- Facebook, Twitter, Instagram, Reddit, YouTube, Amazon

Often outsourced to UX teams – but web developers are involved at almost every step.

- Need to understand and iterate upon feedback
- Necessary for UX analysis (e.g., in assessing observation recording)
- Necessary to support process (e.g., generating different versions of websites)

I'm an aspiring web designer/developer – why do I care?

Some of this may be in the exam.

Data Types

What Can You Expect to Learn?

- A spectrum of data
- Qualitative:
  - Users tell you of problems & situations of which they are aware
  - You observe situations that users may not be fully aware of, due to their immersion
- Quantitative:
  - Measure task performance with existing tools/methods: speed, errors, dead-ends, learning curves for novice users,
  - Numerical data from questionnaires or observation: # of computers owned, # of email messages received per day, # times confused

What Can You Expect to Learn?

- A spectrum of data
- Qualitative:
  - SUBJECTIVE • Users tell you of problems & situations of which they are aware
  - OBJECTIVE • You observe situations that users may not be fully aware of, due to their immersion
- Quantitative:
  - OBJECTIVE • Measure task performance with existing tools/methods: speed, errors, dead-ends, learning curves for novice users,
  - SUBJECTIVE • Numerical data from questionnaires or observation: # of computers owned, # of email messages received per day, # times confused
- For both, what you get is influenced by how you ask the question!

Types of Data

- Data from a survey question asking users to rate their agreement with the statement,
  - "*This technique was easy to learn*" on the following scale:
  - strongly agree, agree, neutral, disagree, strongly disagree
- Data from a survey question asking users to choose their preferred operating system from the choices of Ubuntu, macOS, and Windows
- Quotes from an interview study
- Movement time results from an experiment involving aiming at graphical targets on a computer

Compared: Types of Data

- Quantitative or Qualitative
  - Quantitative
    - Easier to analyze (statistical approaches)
    - Hard to capture subtleties
    - Frequently faster/easier to gather
  - Qualitative
    - More difficult to analyze
    - A lot of rich detail in the data
    - Can be more time-consuming/expensive to gather
- Subjective/Objective
  - Objective
    - Not subject to bias
    - Can be harder to gather
    - Some things cannot be measured objectively
  - Subjective
    - Subject to bias
    - Frequently easier to gather
    - More flexible in terms of things that can be measured

Overview of Methods

Disruptive Methods (Intrusive)

- User observation
- Interviews
- Questionnaires (beliefs/attitudes)
- Observation
- “Think aloud” protocols
- Audio/video recording
- Physiological traces

Non-Disruptive Methods (Non-Intrusive)

- Gaze or eye movement traces
- System logs (including web logs)
- (Hidden) observation
- (Hidden) audio/video recording
- Archives

- Note: non-disruptive ≠ privacy-respecting!

How Do You Choose the Method(s)?

- Depends on goals, questions, & constraints:
  - Control over experiment environment (internal validity), generalizability (external validity) and realism (ecological validity)
  - Natural vs. Artificial setting
  - Objective vs. Interpretive approaches
  - General principles vs. Understanding a specific event
  - Time, cost, expertise, or resources available
  - Stage of development at which the evaluation is performed

Example

- You're designing a new e-commerce site for an existing popular home improvement superstore.
  - What are some activities you might do in the course of a user-centered approach to this design project?
  - When would you do them?
- Hints:
  - Who are the users?
  - What are their tasks and goals?
  - What are their current patterns/contexts of behaviour?

When to Use a Method

- Contextual inquiry
- Ethnography
- Observation methods

Understand your users and/or their task before concept

- Interviews
- Questionnaires & surveys

Any time, any place

- “Discount” (expert evaluator) methods

On a prototype

Methods

Observation

Types of Observations

- Simple observation
- Think-aloud

Simple Observation

- User is given a task (or not), and evaluator just watches the user
- Problem: no insight into the user’s decision process or attitude

Think-Aloud Protocol

- Subjects are asked to say what they are thinking/doing:
  - What they believe is happening
  - What they are trying to do
  - Why they took an action
- Gives insight into what the user is thinking

Think-Aloud Protocol

- Problems:
  - Awkward/uncomfortable for subject (thinking aloud is not normal!)
  - “Thinking” about it may alter the way people perform their task
  - Hard to talk when they are concentrating on problem
- Still is the most widely used method in industry

Retrospective Think-aloud

- Avoid problems of the Think-Aloud Protocol by videotaping the experience and performing a retrospective think-aloud

- Problems:
  - Awkwardness of watching themselves on video
  - Awkwardness of reliving mistakes
  - Reflection of the experience rather than in context

Think-Aloud Protocol

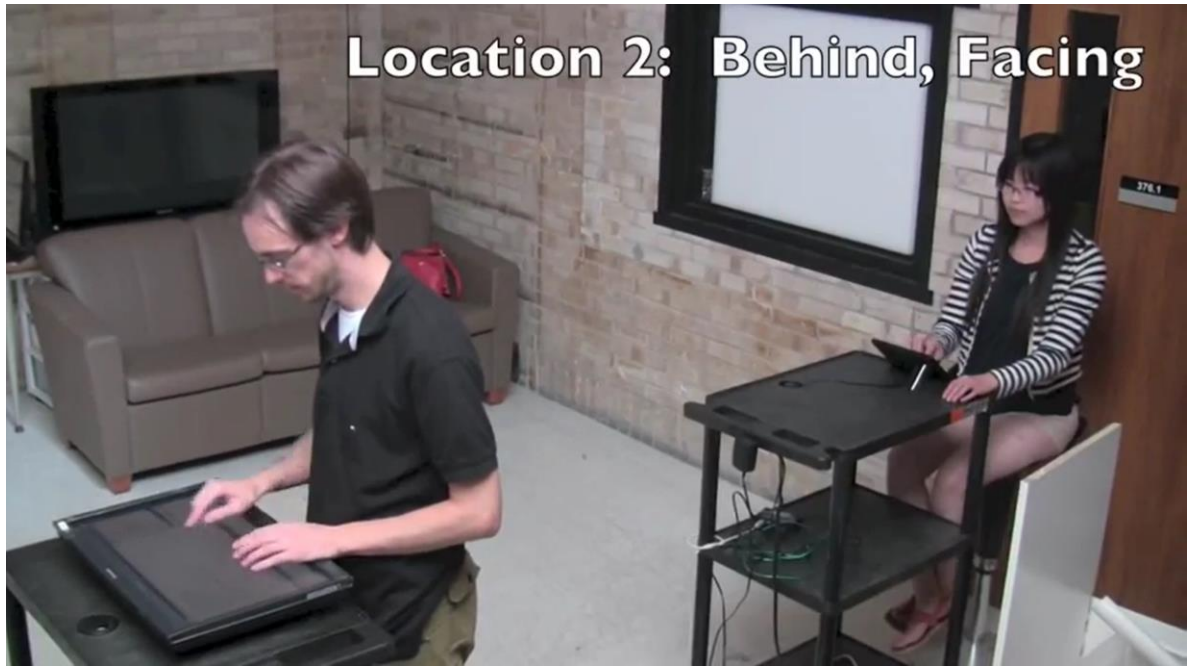
- Example
  - Eishita, F. Z., Stanley, K. G., & Mandryk, R. (2014, October). Iterative design of an augmented reality game and level-editing tool for use in the classroom. In *Games Media Entertainment (GEM), 2014 IEEE* (pp. 1-4). IEEE. https://doi.org/10.1109/GEM.2014.7048093

  - Direct observation technique used in iterative design process for a classroom tool

Group Name	Editor	High Level Observation
Beta Version School Groups Grade 7	• Provided teams with specific subtopics to help focus learning and speed design • Participants prepared 3 levels of paper prototype and picked the best	• The 'guess the commonality' method of play was difficult to learn for the age range • The players did not feel the rush to finish the game since there was no set time limit.
Camp Week1 Grade 7,8	• We continued providing teams with specific subtopics from wiki page • Participants prepared 3 levels of paper prototype and picked the best t	• Children learned the game in shorter time • Timed game increased the pace and excitement
Camp Week2 Grade 5,6	• Preparing levels with provided subtopic continued	• The learning time was higher with younger students. • Smartphone manipulation was difficult for some
Camp Week3 Grade 7,8	• We asked the players to prepare and implement only 1 scramble	• Learning time continued was similar
Camp Week4 Grade 5,6	• Preparing 1 clue with provided topic continued	• Players were moving actively • Preparing one clue made it interesting and easier
Camp Week5 Grade 7,8	• Preparing 1 clue with provided topic continued	• Learning time continued to be similar to the previous levels • 'Fill in the blank' question pattern was more popular

Recording Observations

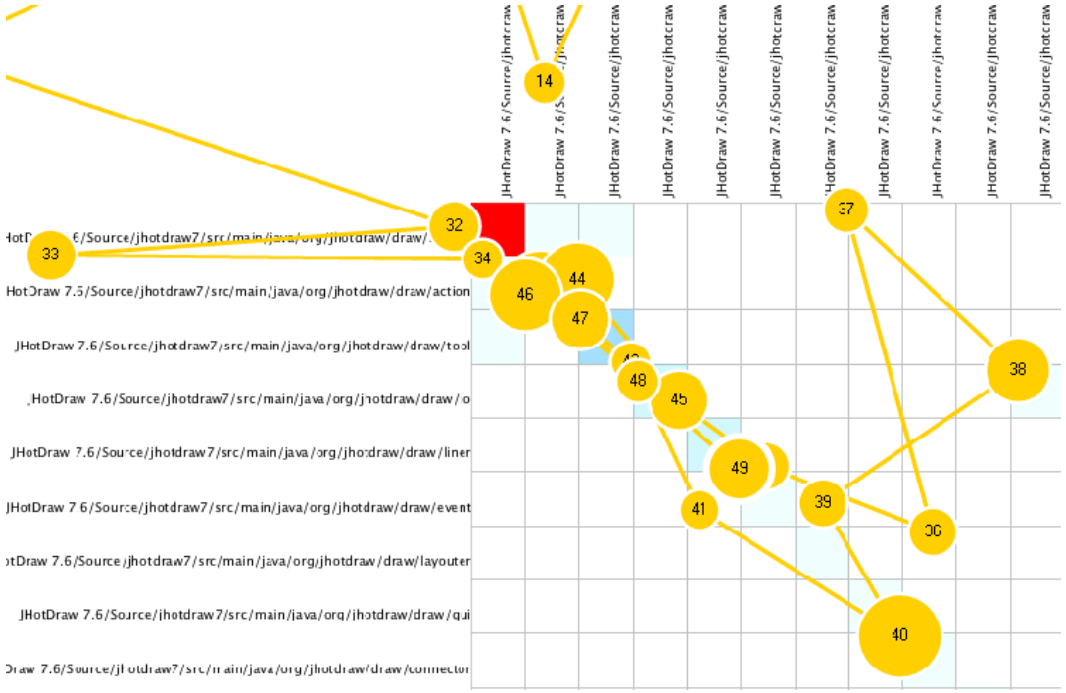
- Companies often build “usability labs” with one-way mirrors, video cams, etc.



Eye Tracking

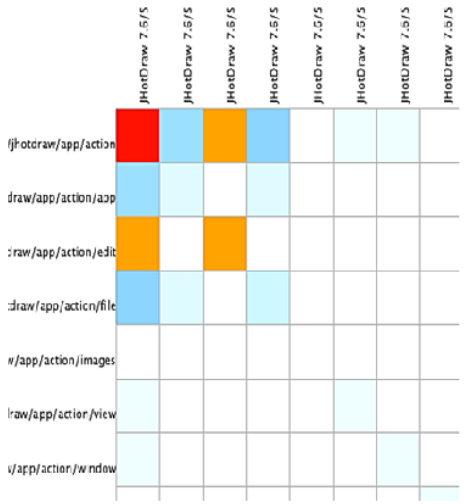
Eye Tracking

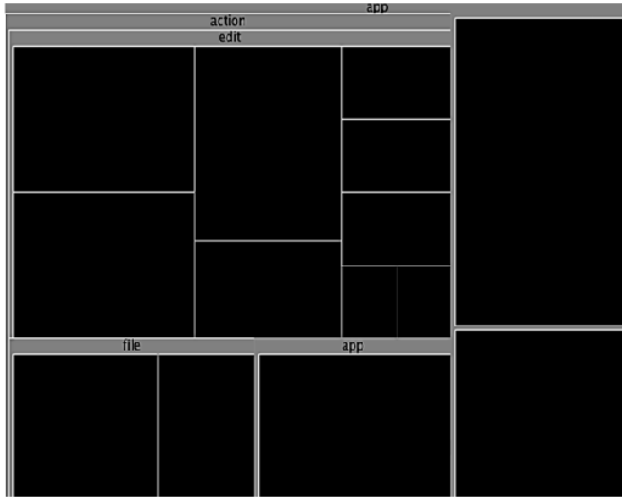
- Shows areas where users focused their visual attention



Eye Tracking

- Example
  - Uddin, M., Gaur, V., Gutwin, C., Roy, C. 2015. On the Comprehension of Code Clone Visualizations: A Controlled Study using Eye Tracking. In 15th IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM 15), Bremen, Germany. 161-170. https://dx.doi.org/10.1109/SCAM.2015.7335412

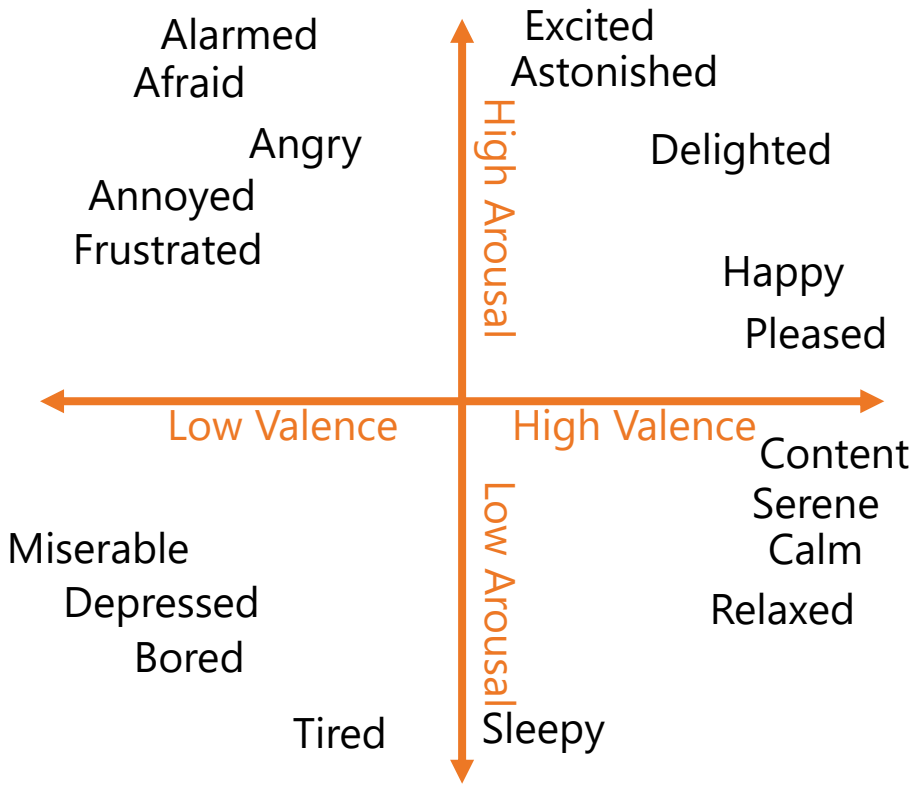




  - Testing code clone visualizations

Pupil Diameter

- May indicate mental workload or valence
- Difficult to apply due to sensitivity to light



Logging

Logging

- System generated events and times
- Works for websites, apps, and games

Logging Tools

- Simple text file
- Database
- Commercial logging tools (e.g. Hockeyapp)

What to Log?

- Spectrum of possibilities from high level to low level
  - System Engagement
  - Task completion
  - Navigation
  - Events
  - Complete input logging/video capture

What to Log?

- System Engagement
  - Measure usage at a high level
    - E.g. Activations/launches per day, total time used per day, etc
  - Works well for long-term studies, doesn't generally apply to lab studies

  - Benefits:
    - simple overall measure of user response to system
  - Drawbacks:
    - hard to understand the reason for a given result

What to Log?

- Task Completion
  - Measure user behavior in terms of tasks/task outcome
    - E.g. successfully navigated to 3 destinations, failed to navigate to 2
  - Benefits:
    - Simple per-task measurement of system effectiveness
    - Can get some indication of why failures occur based on progress into task flow
  - Drawbacks:
    - Some systems don’t have a clear definition of task or usage cannot be broken down into a reasonable number of tasks easily (e.g. Microsoft Word)
    - May not know specific reasons for task failures

What to Log?

- Navigation
  - Record user interaction as a series of ‘page views’
    - E.g. home > search > result detail > cart
  - Benefits:
    - Good level of detail without overwhelming amount of data
    - Can see where users are getting stuck or quitting
  - Drawbacks:
    - Might not be meaningful if system has few screens/views
    - Might know what page something is going wrong on, but not why

What to Log?

- Events
  - Log every time anything happens within the system
  - “Anything” can be very low level:
    - E.g. Press Button {name: settings}, Press Button {name: advanced}, Press Checkbox {name: show full names, state: on}, Press Button {name: save},
  - Or higher level
    - E.g. Navigate {page: settings}, Change setting {name: show full names, value: on}
  - Benefits:
    - Detailed understanding of interactions
  - Drawbacks:
    - Time consuming to analyze, especially if lower level
    - Might forget to log certain events, especially if high level

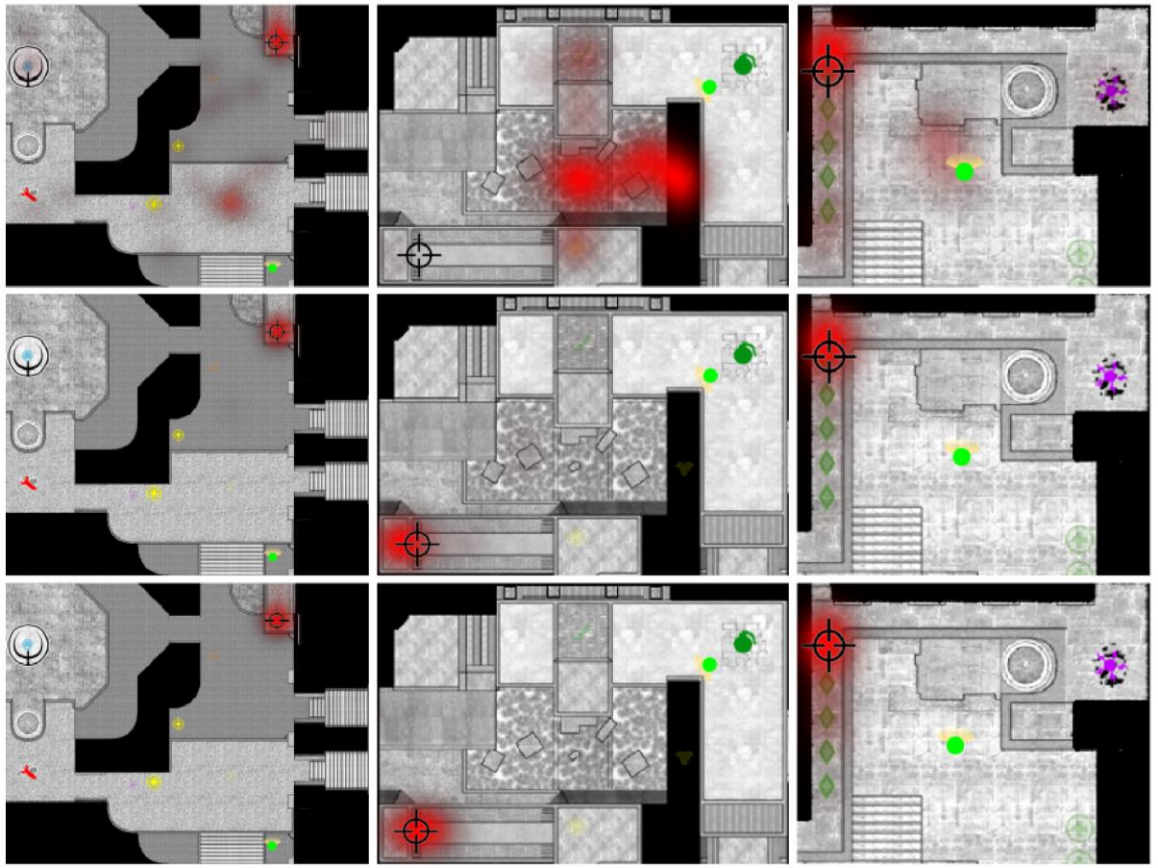
What to Log?

- Complete input logging/video capture
  - Log all input to the system
    - Mouse coordinates every time the cursor moves
    - All key/button presses
    - Other inputs (joystick, touch, etc)
  - Enables researchers to completely replicate the interaction at a later time
    - If the system has randomness or timing as factors in it's operations, this won't be possible; a video screen capture is helpful in this case
  - Good as a "back-up" in case other logging fails or additional data is needed for unanticipated analysis

  - Benefits:
    - Complete understanding of interactions
  - Drawbacks:
    - Very time consuming to analyze or interpret

Example

- Example
  - Johanson, C., & Mandryk, R. L. (2016, May). Scaffolding player location awareness through audio cues in first-person shooters. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 3450-3461). ACM. https://doi.org/10.1145/2858036.2858172
- Logged player position at all times in the games
- Generated heat maps after the fact to help with data analysis



Interviews

Querying Users Via Interviews

- "Conversations with a purpose"
- Excellent for pursuing specific issues
  - More interaction than with observation:
    address specific questions of interest
  - More flexible than questionnaires:
    probe more deeply on interesting issues as they arise

Querying Users Via Interviews

- “Conversations with a purpose”
- Problems
  - Accounts are subjective
  - Time consuming (to conduct and to analyze)
  - Evaluator can bias the interview
  - Prone to rationalization of events/thoughts by user
    - User’s reconstruction may be wrong

Retrospective Interview

- Post-test interview to clarify events that occurred during system use:
- Record what happened, replay it, and ask about it

- Pros:
  - Excellent for grounding a post-test interview
  - Avoids erroneous reconstruction
  - Users often offer concrete suggestions
- Cons:
  - Requires a second session

Things You Uncover During Interviews

- Exceptions
  - Lots of things people do are not “in the manual”
  - Many jobs evolve to fit changing circumstances
  - Much of this is not documented
  - Many times “management” does not know about this

Things You Uncover During Interviews

- Domain knowledge
  - Most people know a lot about their jobs, and those they work with
- Terminology, common phrases, specific details
  - Audio recording helps capture this
  - Video recording helps provide body language
  - Written notes can provide context, but not always details

Questionnaires (Surveys)

Querying Users Via Questionnaires

- Closed or open questions
- Evidence of wide general opinion
- Only as good as the questions asked

- Pros:
  - Preparation “expensive,” but administration cheap
    - Can reach a wide subject group (e.g., mail or email)
  - Does not require presence of evaluator
  - Results can be quantified
- Cons:
  - Can have low response rate and/or low quality response

Styles of Questions: Open-ended

- Asks for opinions
- Good for general subjective information
  - But difficult to analyze rigorously

- Examples:
  - Can you suggest any improvements to the interface?
  - What did you like about the system?

Styles of Questions: Closed

- Restricts responses by supplying the choices for answers
- Can be easily analyzed …
- But can still be hard to interpret, if questions / responses not well designed!
  - Alternative answers should be very specific

- Examples:
  - Do you use computers at work:
    - ○ often
    - ● sometimes
    - ○ rarely

  - In your typical work day, do you use computers:
    - ○ over 4 hrs a day
    - ○ between 2 and 4 hrs daily
    - ● between 1 and 2 hrs daily
    - ○ less than 1 hr a day

Styles of Questions: Scalar - Likert Scale

- Measure opinions, attitudes, and beliefs
- Ask user to judge a specific statement on a numeric scale
- Scale usually corresponds to agreement or disagreement with a statement

- Example:
  - Characters on the computer screen are hard to read:
    - strongly agree 1 ○ 2 ○ 3 ● 4 ○ 5 ○ strongly disagree

Styles of Questions: Multi-Choice

- Respondent offered a choice of explicit responses

- Examples:
  - How do you most often get help with the system? (tick one)
    - ○ online manual
    - ● paper manual
    - ○ ask a colleague

    Which types of software have you used? (tick all that apply)
    - ● word processor
    - ○ database
    - ● spreadsheet
    - ○ compiler

Styles of Questions: Ranked

- Respondent places an ordering on items in a list
  - If done digitally, often uses drag & drop to reorder items
- Useful to indicate a user's preferences
- Forced choice

- Example:
  - Rank the usefulness of these methods of issuing a command (1 being most useful, 2 next most useful..., 0 if not used)
    - _2_ command line
    - _1_ menu selection
    - _3_ control key accelerator

Styles of Questions: Combining Open-Ended & Closed Questions

- Gets specific response, but allows room for user's opinion

- Example:
  - Characters on the computer screen are hard to read:
    - Strongly agree 1 ○ 2 ○ 3 ● 4 ○ 5 ○ Strongly disagree
    - Comment: *the undo function only works part of the time*

Drawbacks of surveys

- Capture attitudes, not behaviours
  - Possibly self-deception, possibly unawareness
  - In general, people are just pretty bad at remembering their behaviours accurately
- “Everybody lies”
  - The truth about lying in online dating profiles
    - http://portal.acm.org/citation.cfm?id=1240624.1240697
    - Sometimes we don’t even know our own biases.
      - (Implicit association test) IAT

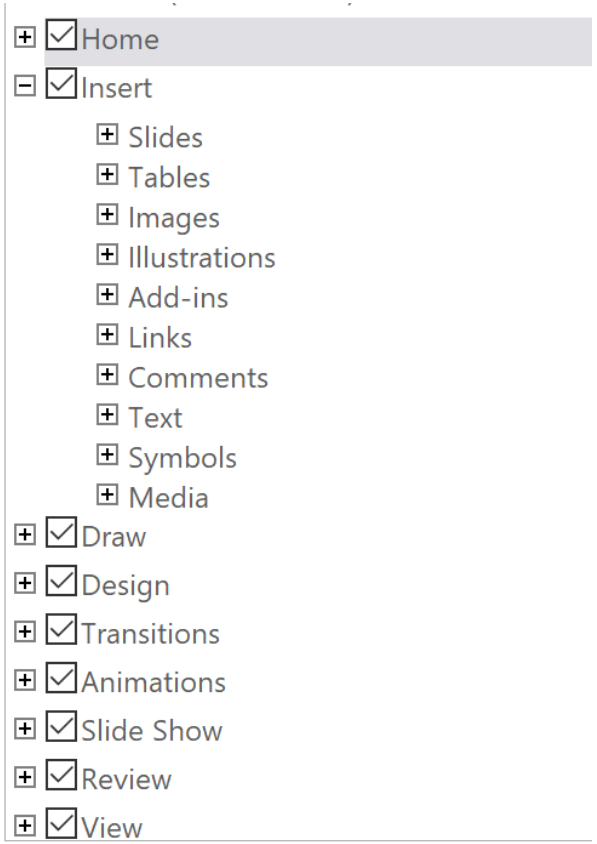
Tree Testing

Tree Testing

- Also called reverse card sorting or card-based classification
- Used to test the efficacy of an existing navigation or menu hierarchy

Tree Testing - Method

- Extract the tree structure out of the menu system or website
- Users receive a set of tasks or items to find
- For each task/item, participants browse the tree structure until they find the category they believe is correct
- Participants' guesses are recorded
  - Each category which is expanded or browsed before the final selection may be recorded as well
  - Timings may also be recorded



Tree Testing - Analysis

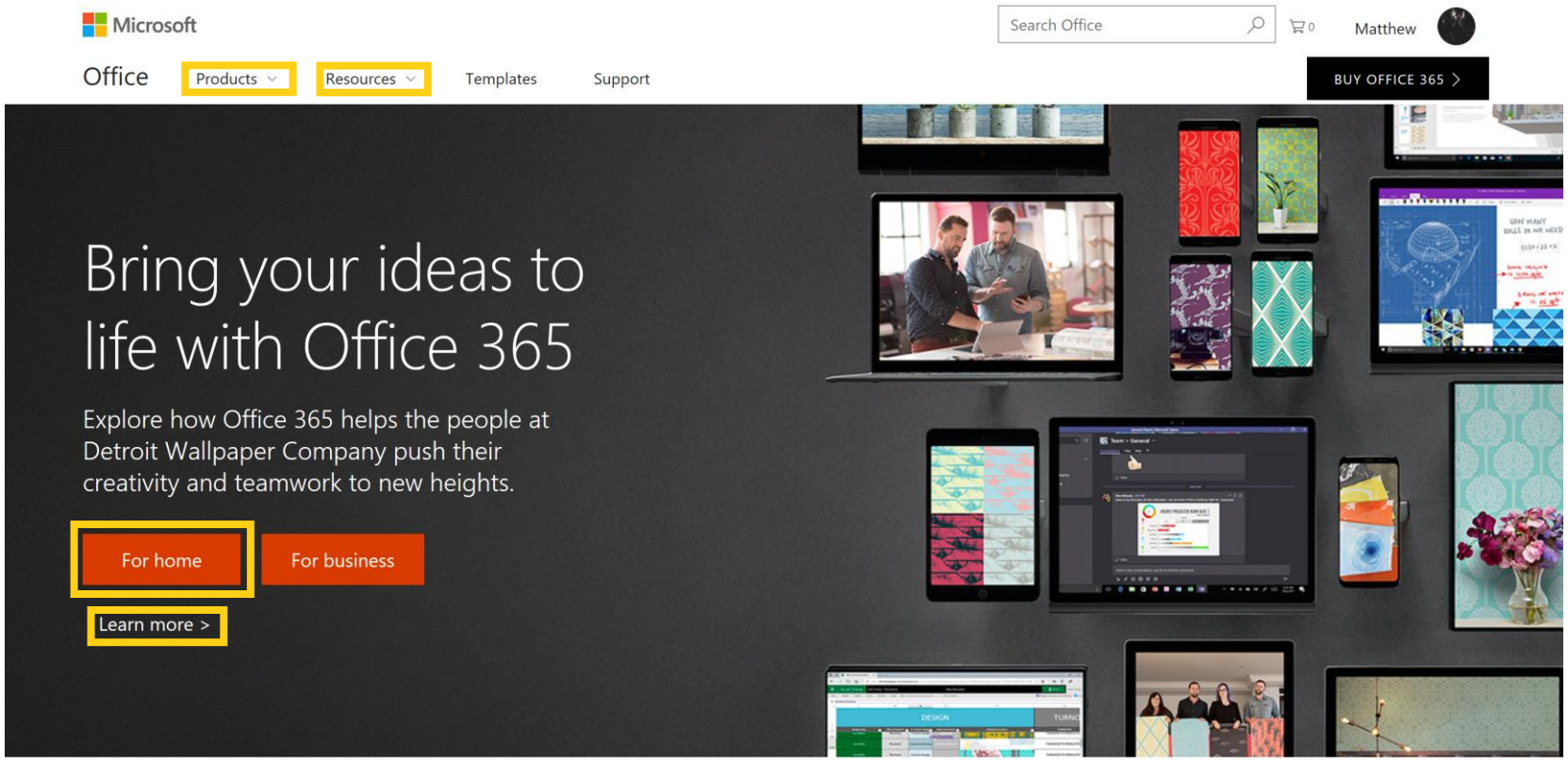
- Main analysis consists or measuring the number of correct responses for each task or item
- Variety or secondary analysis possible:
  - For each task/item, which incorrect categories were chosen most frequently
  - Which categories result in the most incorrect guesses
  - Which tasks/items cause the user to explore more before making a final decision
  - Which tasks/items took longer to find an answer for
  - Others…

First Click Testing

First Click Testing

- A method for determining the success of navigation in a system
- Users are asked to consider different tasks
- For each task, users click a single item in the UI
- This click is evaluated as right or wrong based on the necessary actions to achieve the task
- Studies show that people who make the correct first click end up completing a task 87% of the time while those that make the wrong first click end up completing a task only 46% of the time.

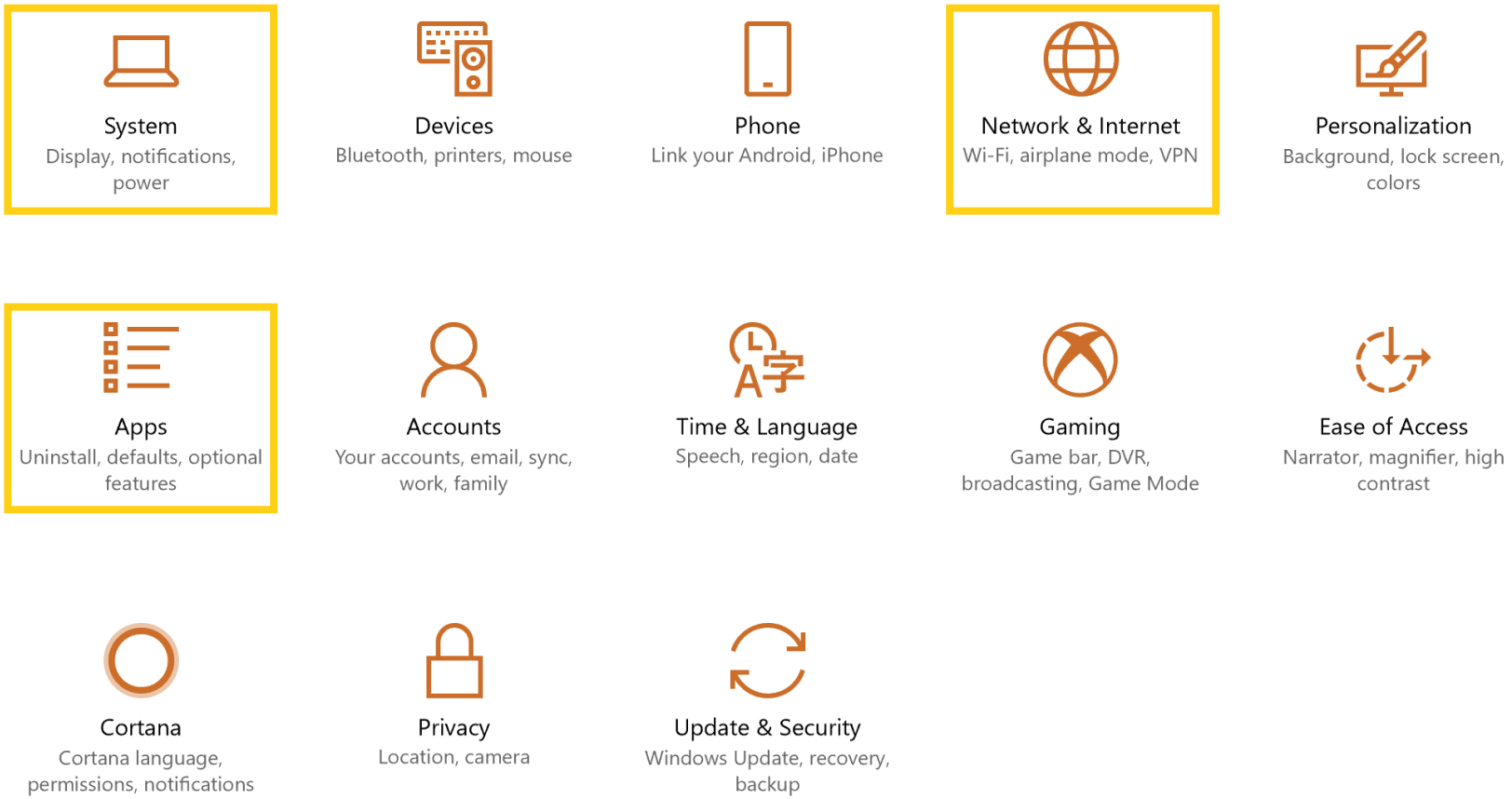
https://www.usability.gov/how-to-and-tools/methods/first-click-testing.html



What is Office 365

Learn more about Office 365

https://www.microsoft.com/

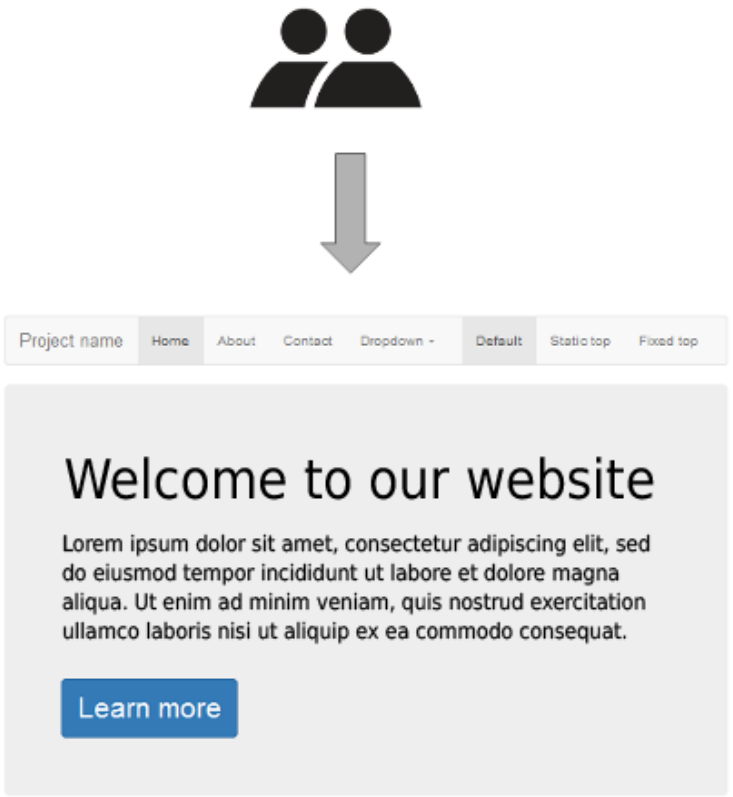

System
Display, notifications, power
Devices
Bluetooth, printers, mouse
Phone
Link your Android, iPhone
Network & Internet
Wi-Fi, airplane mode, VPN
Personalization
Background, lock screen, colors
Apps
Uninstall, defaults, optional features
Accounts
Your accounts, email, sync, work, family
Time & Language
Speech, region, date
Gaming
Game bar, DVR, broadcasting, Game Mode
Ease of Access
Narrator, magnifier, high contrast
Cortana
Cortana language, permissions, notifications
Privacy
Location, camera
Update & Security
Windows Update, recovery, backup

A/B Testing

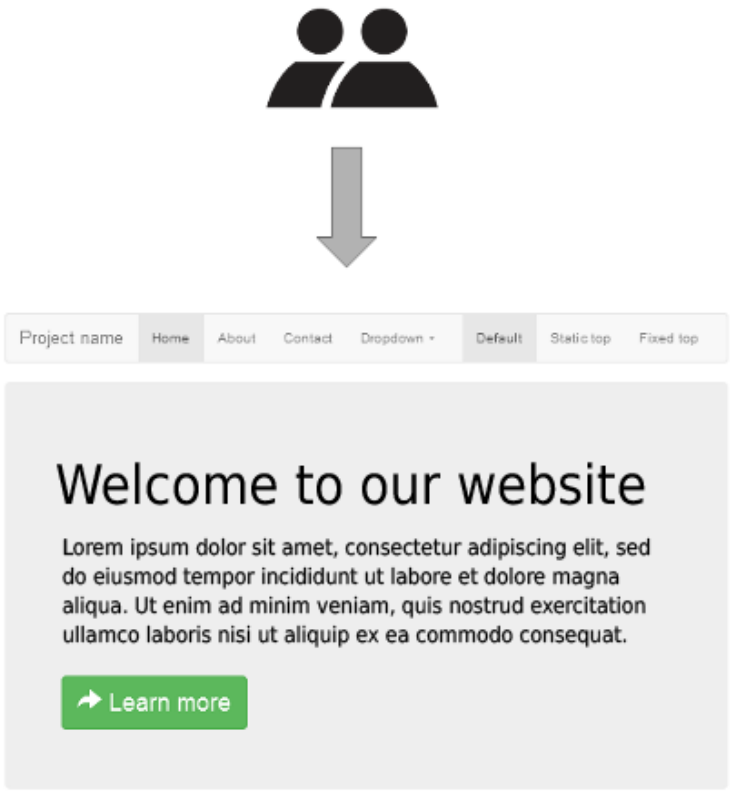
A/B Testing

- Also known as ‘split testing’
- Roll out two different versions of the same website – can be done live
- Use previous methods described (e.g., questionnaire, logs) to assess participant response, evaluation, and engagement

A/B Testing



Click rate: 52 %



72 %

A/B Testing

- You've likely been a participant in one of these tests yourself ☺
- Facebook, Twitter, Instagram – all major social media platforms have done this
- Also often used by advertising companies – which ads generate the highest clickthrough rates?

Summary

- There are multiple ways of involving users
- There are many types of data you can collect
  - Pick an appropriate one!

P.S.

If you find this topic particularly interesting – consider a career in UXR (User Experience Research)!