ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

HUMAN COMPUTER INTERACTION GROUP

HCI Evaluation Techniques

School of Computer and Communication Sciences

EPFL

Pearl Pu

STUDENTS' MASTER THESIS WORK

- "This GUI app design was a very challenging one: Most user will use this application in stressful situation, where sometimes patient's life is at sake, thus the GUI should be simple and efficient. Also, each medical service has a different way of planning its shifts, thus the shift editor GUI had to be powerful enough.
- As I told you today, x's work has been greatly appreciated at HUG. The application she has delivered is going to go into production in June . The users are very happy about it, and so is the IT department, as it helped set new technological standards for our next applications, and showed us a brand new way of thinking about the GUI design, focusing on user goals, and not only functionalities."

- March 31, 2014 Hôpitaux Universitaires de Genève Direction des systèmes d'information

What is evaluation in HCI?

- How to evaluate software other than HCI
  - Satisfies design spec, bug-free, run-time efficiency, memory usage, parallelism, reliability, robustness, susceptibility to attacks etc.

- How to evaluate the HCI aspect of a software?
  - Evaluate interaction behavior, visual interface design, content
  - How?
  - There are many methods

- Revisit the three criteria of HCI
  - perceived usefulness of software
  - ease of use
  - playfulness

WHY EVALUATION IS HARD?

- Developing a measuring standard for usability is hard
  - user errors and task time?
  - what type of tasks & errors? more refined criteria?
  - measuring satisfaction is a multi-criteria decision
  - some researchers dedicate their careers to understand and develop user adoption principles (technology adoption model)

- In evaluation, our goal is also to balance subjectivity and objectivity

A RANGE OF METHODS

- Focus Group
- Walk-throughs
- Survey
- Open and Closed Card Sorting
- Expert
- Heuristic Evaluations
- Formative Study
- Comparative Testing (A/B)

FOUR HCI EVALUATION TECHNIQUES

- Critiquing
- Inspection method
- Heuristic evaluation
- Comparative user study (A/B)

Critiquing

Definition of critiquing

- ***Critiquing*** is a process where an HCI designer or a usability expert identifies a usability problem or an area of usability problems using established ***design principles***

CRITIQUING

How to do a critiquing

1. Locate the problem
2. Explain the problem using HCI language
3. In what context the problem is most severe
4. Propose fixes

Catalog of design principles has been developed by HCI@EPFL
(see resource in moodle)

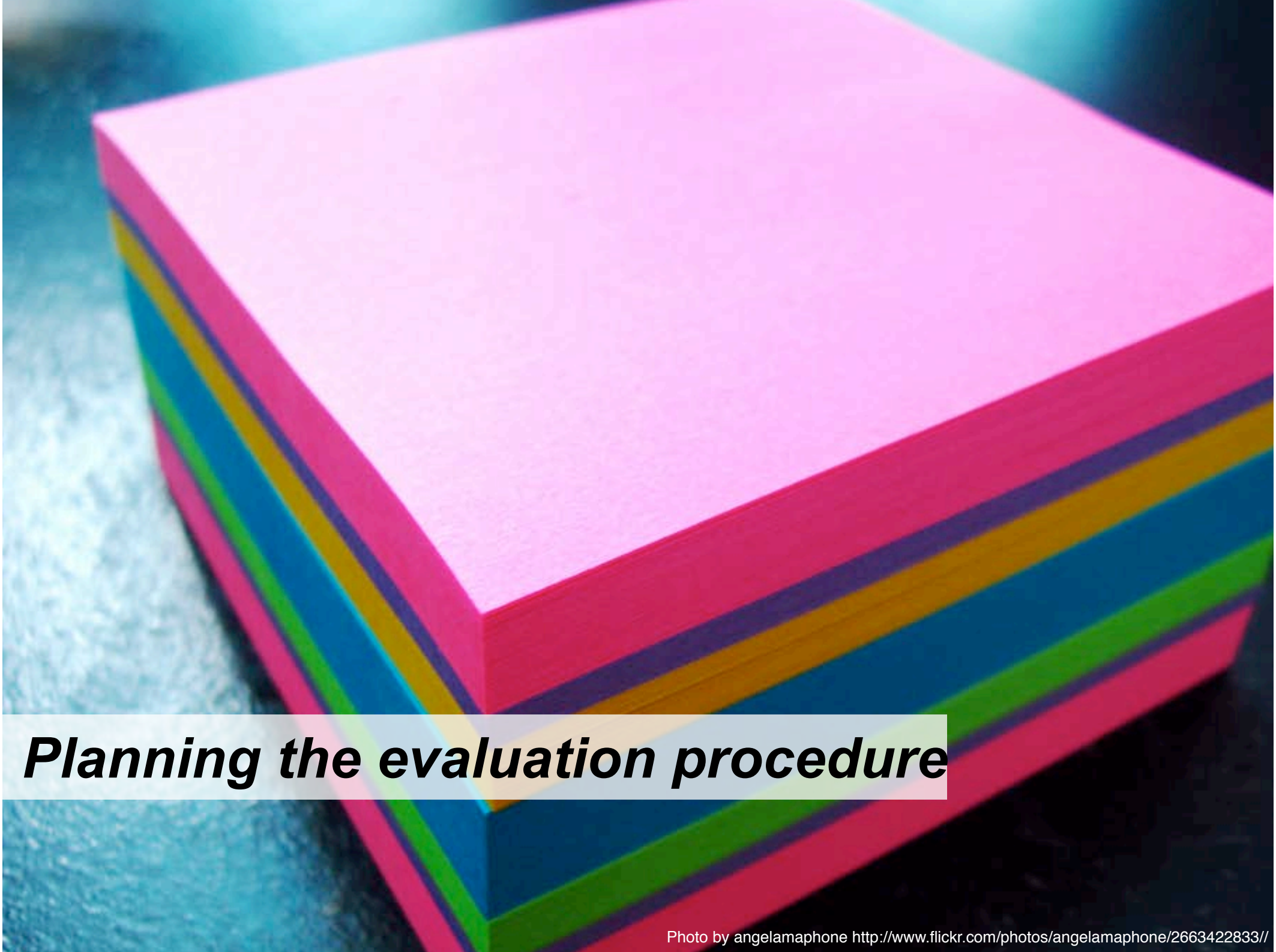
Inspection

Definition of inspection

- ***Inspection*** is a process where usability experts evaluate with ***users*** on a specific area of the system

INSPECTION

Identify usability problems by evaluating with users on a specific area of the system



Planning the evaluation procedure

Photo by angelamaphone http://www.flickr.com/photos/angelamaphone/2663422833//

What is a typical procedure?

5 steps

1. Analyze the user segments
2. Design user tasks to reveal the usability problems
3. Decide on how to contact the users
4. Prepare scripts to inform users of the interview
5. Design a post-study questionnaire

Step 1: Analyzing the user segments


What different sorts of people might use this product?

1) Analyzing the user segments

- Who are the main users of this system?
  - Old days: demographics
  - Nowadays:
    - Expert
    - Intermediates
    - Beginners
  - Who? prefer beginners and intermediates
  - Who? experts for dedicated systems

Step 2: design user tasks

- Design tasks that correspond to the main components of the Task Tree

- **Task 1:** Make a flight reservation from Geneva to London for the dates November 10-20, pay, and print
- **Task 2:** Change the dates and make a reservation for November 12-22
- **Task 3:** Retrieve an old trip, modify its dates, and book it again

How to collect evidence of user frustrations?

- Measure task completion time, mouse clicks and number of errors
  - Method: take notes, screen recording
  - Quantitative feedback

- Observe the users while they are doing the tasks
  - Method: Ask the users to think-aloud, take notes, record video (audio and non-verbal feedback)
  - Qualitative feedback

Step 3: how to recruit users

- Who?
  - friends, friends of friends, friends of family members
  - colleagues
  - university staff, professors, instructors
  - 2nd level relationships are better
  - center of testing (best)
- How to recruit them
  - email, social media, professional contacts
  - subscription to testing centers
  - Consider small gifts to show appreciation

Step 4: scripts to inform users

- Write in advance scripts related to the test procedure
  - Scripts for greeting users, task instructions
  - Make users aware of the procedure
  - What is the purpose, how long it will take, where/ when it will happen?
  - Is there any privacy considerations?
    - e.g. video recordings will only be used for this project

Ensure all participants are given the same instructions

Example orientation script

The goal of our project is to evaluate the online flight reservation website «Flight Booking». The results of our evaluation will be summarized as part of a semester project for the Human-Computer Interaction course given at EPFL.

You will be asked to use the website and do certain tasks. While doing the tasks, you will be asked to «think aloud». What you say will be noted and also video recorded for further analysis. Your video will only be used for this project and will not be distributed further. We will also record your computer screen.

Please review and sign the attached consent form to confirm your participation. The fact that you sign the consent form does not mean that you can not cancel your participation. Please let us know if you have any questions.

Thank you very much for participating in this study.

EXAMPLE INTERVIEW PROCEDURE SCRIPT

The purpose of our work today is to explore «Flight Booking» website and to identify usability features that could be improved.

I will first give you 3 tasks to do. We will record you while you are doing these tasks. Please keep in mind that it is the website that is evaluated here, not yourself. Feel free to work on each task at a pace that is normal and comfortable for you.

Think aloud as you do the tasks to describe what you are doing. I will remind you from time-to-time if you forget to think aloud.

After the completion of each task, I will interview you to learn more about your experience with the system.

Before we proceed, do you have any questions?

4) Post-study interview

- Ask open-ended questions about the overall experience
  - What are your overall impressions of the website?
  - What are the difficulties you encountered?
  - If you want to make one significant change on the website, what would it be?
  - Would you recommend this site to other people?

Conducting the interview

What is a typical procedure?

- Make at least one pilot-study to ensure the protocol is working
  - Update if necessary
- Decide who takes the facilitator and observer/noter role
  - At least two people is needed
- Set-up and test the recording equipment before each interview
- Greet the user and talk about the test procedure
- Ask them to perform the tasks while observer is taking notes and making recordings
- Do the post-interview session

Photo by Dave Gilbert http://www.flickr.com/photos/eye2eye/50892860/

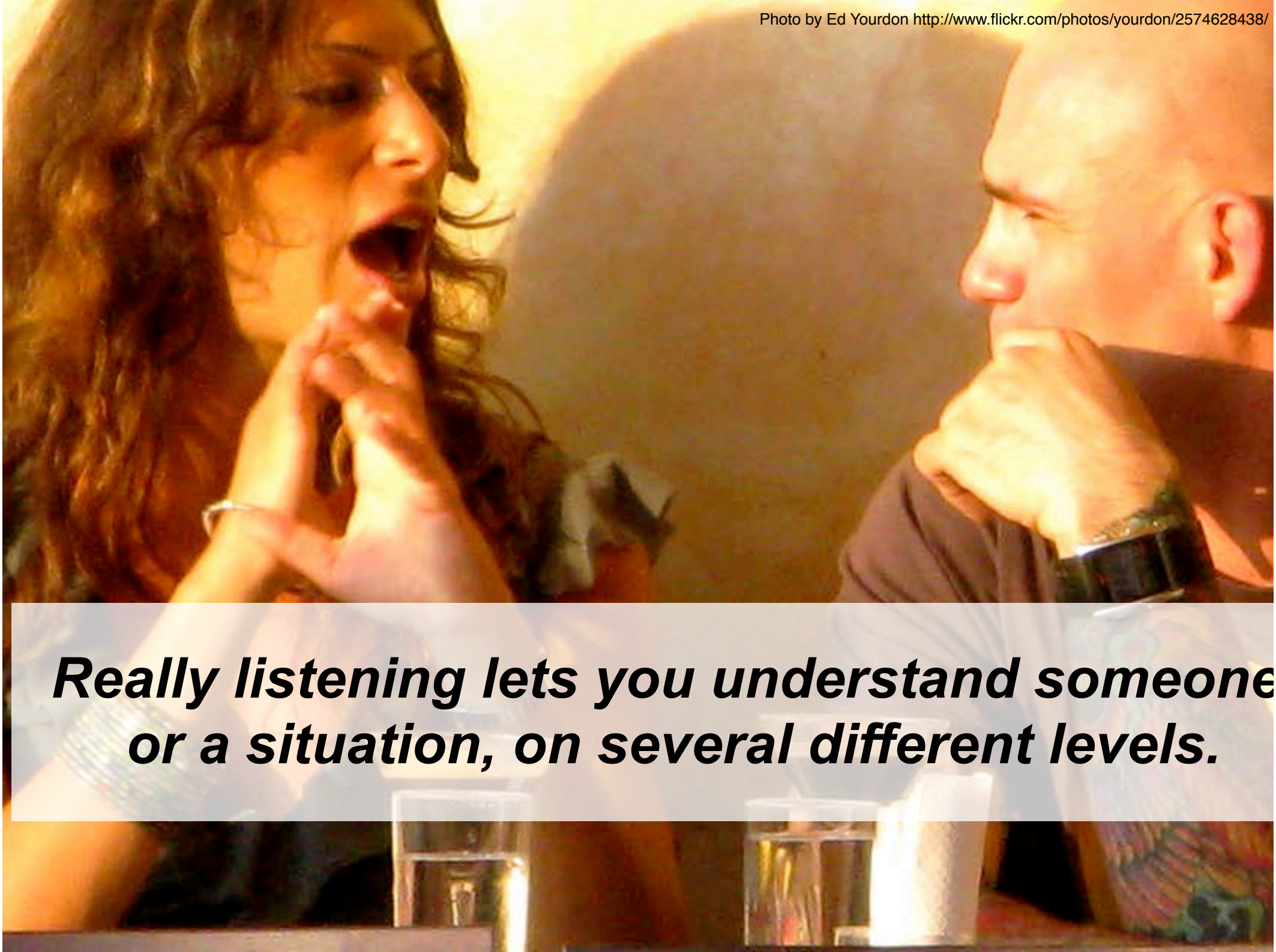


Photo by Ed Yourdon http://www.flickr.com/photos/yourdon/2574628438/
Really listening lets you understand someone
or a situation, on several different levels.

Active Listening

e.g. “Tell me more about that…”

Receive

Appreciate

Summarize

Ask

From Julian Treasure’s TED talk ‘5 ways to listen better’

Listening to body languages

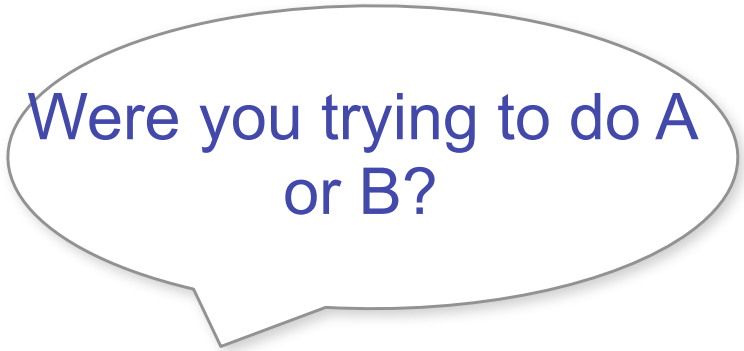
Ask open questions – don't lead

YAY

- Who
- What
- When
- Where
- Why
- How

NAY

- Did
- Have
- Are
- Were
- Will



How to keep people talking

Tell me more…

What do you mean by…

What else can you tell me about…

Can you tell me the story about that?

Help me understand better

Don't Judge

e.g. You wouldn't go on Pirates of the Caribbean?!

http://www.flickr.com/photos/18796746@N05/4272817915/

How to deal with difficult people

Photo by David Anderson http://www.flickr.com/photos/venndiagram/4667350842/

- Encourage them to say more.
- How does this (part/page) compare with what you were expecting?
- Could you tell me more about your story on using this device?
- What did you do when the problem happened?

- Bring them back to the track immediately and politely.
- That's really interesting, thanks for telling me about that. To come back to....
- Can I interrupt you? Sorry, I was actually curious if you could...

Synthesizing the results

What is a typical procedure?

- Go over the screen recordings and make a summary of the task measurements
  - Use tables, charts
- Go over the notes taken, video/audio recordings and answers to post-study questionnaire
  - Discuss them within the team
  - Note down important points
- Revisit the usability problems and find evidences of user frustrations to support each of them
- Think about whether there are differences between your initial ideas and what you have learned from the interview
- Consider what you learned from the interview can guide the re-design of the system
- Report the results in a professional way

Quantitative Analysis

- Organize the data for each user in spreadsheets
- Simple statistical analysis
  - Averages (mean, median, mode)
  - Use summary charts and tables

	Task completion time	Number of clicks	Number of errors
Task 1	49	15	3
Task 2	56	20	5
Task 3	35	6	2

- Use the results as an evidence of usability problems
  - All the users were frustrated while doing Task 1. On avg, they made 3 errors.
  - While novice users made more errors on Task 1, both type of users made errors on Task 2.
  - …

QUALITATIVE ANALYSIS

- Transcribe think-aloud protocols and pay attention to non-verbal feedback
  - Find excerpts that supports user frustrations
    - “Normally, it should let me choose the dates from the calendar, what should be the format of the date?” **(facial shrug)**
    - “Do I have to enter all this information from the beginning?” **(bored face)**
    - “I am not able to take a print-out, did they send it to my email address?” **(sarcastic laughing)**
- Transcribe the answer to post-study interview
  - Find excerpts that supports user frustrations, learn from the user
    - “I was quiet surprised when the website didn’t automatically complete my source and destination airports and didn’t provide me a pop-up calendar.
    - “I think the most annoying aspect is the requirement to enter all my information from scratch when I change the travel dates”

Discussion of results

- Make a sound and realistic discussion on what you hypothesized and what you have learned
- Mention any useful findings that can guide the re-design of the system
- Example:
  - «We were able to find evidences of user frustrations for each of the usability problems that we initially suspected. What is more, we have learned that … One comment from the expert user shed light to how to fix the problem on...

SUMMARY ON INSPECTION

1. Make a hypothesis
2. Choose an area of a system based on hypotheses
   - e.g., Flight Booking
   - we can inspect booking, booking modification, choosing options from a list, check-out, payment, printing, etc.
3. Design tasks that users are most likely to perform in those areas
4. Ask users to perform these tasks
5. Ask them what they think of that aspect of the system

How to combine critiquing with inspection

- You can suspect a usability problem using critiquing
- To ensure objectivity, you then design an inspection procedure to perform a testing with a task script
- Since it involves users, the procedure looks a bit like “qualitative research”

Problems found in Flight Booking

- Usability problem 1: The system does not have auto-completion functionality for the source/destination cities and pop-up calendar for date entry.
- Usability problem 2: The system does not remember user's information when the user wants to change dates
- Usability problem 3: The system does not let the user print the reservation

Heuristic Evaluation Methods

HEURISTIC EVALUATION METHOD

- HCI experts identify usability problems following Nielsen's ***10 principles***
- These principles are called "heuristics"

http://www.useit.com/papers/heuristic/

Heuristic Evaluation

- 10 heuristics first invented by Nielsen and Molich 1990.
- This method is easy and quick to follow; that's why it's called "discount" usability evaluation
- HCI@EPFL added several interpretations of the heuristics to the list from our own experience

http://www.useit.com/papers/heuristic/

Heuristic 1: Visibility of System Status

The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

- Put yourself in the users' shoes and ask:
- "Where am I in my task?"
- "Where can I go next?"
- "For each action I want to do next, can I find the UI?"
- ***"For each UI, can I predict the system's reaction before I click on it?"***
- "Is there appropriate feedback within a reasonable time following each user interaction?"

H2: Match between system and the real world

The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

- users' mental model and represented model: do they match?
- user mental model = task tree (or context scenario)
- represented model = the software interaction behavior you are responsible for

H3: User Control and Freedom

Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

- Can users back out of actions?
- Complement to this principle, we add:
  - Are users unknowingly given “intelligent software”?
  - Second guessing users’ intentions is really hard; make sure we get it right, or it backfires
  - To balance intelligent software with user control, it’s important to give them a way to opt out
  - E.g., the filter bubble story: people want to read news of opposite opinions, not just tailored to what they always read

H4: Consistency and Standards

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

- This principle means:
  - Use of language consistency
  - Visual and layout consistency
  - Screen consistency
- The same word, situation, or action should always have same meaning or effect
- We add: external consistency, which is consistency with convention used elsewhere

H5: ERROR PREVENTION

Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.

- Error prevention
  - prevent problems from occurring in the first place (e.g., let users select file from menu rather than asking them to type the file names; confirm with users before dangerous actions; avoid using similar command names)
- We add: can users store intermediary results?
  - For major tasks, can users store intermediary results to avoid loosing them?

H6: MEMORY

Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

- Memory (Recognition rather than recall) -- make objects, actions, and options visible; provide examples, default values, easily retrievable instructions

H7: Flexibility & efficiency

Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

- Accelerators -- abbreviations, command keys, type-ahead, edit and reissue previous commands, menu of most recently used files, macros
- We add: allow users to tailor the use of frequent actions to their own key assignments

H8: Aesthetic and minimalist design

Dialogues (or interfaces) should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

- Visual layout principles
  - group elements on an interface
  - structure elements on an interface
  - proximity & alignment principles
  - visual hierarchy
  - use appropriate fonts and colors

Less is more

H9: Help users recover from errors

Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

- Good error message -- error messages should be expressed in plain language (no codes), do not offend users, precisely indicate problem, and constructively suggest solution.
- Related to H3 (control and freedom) Give clearly marked exits -- provide clearly marked “emergency exit” to leave mistakenly entered state (e.g., cancel, undo, redo)

H10: HELP DOCUMENTS

Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

- Help and documentation -- easy to search, focused on user task, list concrete steps, liberal use of examples, not overwhelmingly large.
- For our course project, we will not create help documents.

Comparative Evaluation

COMPARATIVE EVALUATION

- A ***comparative evaluation*** is a user study involving comparing two interactive systems and analyzing the effects of these systems have on users.
- Most quantitative evaluation

EXPERIMENT FORMAT

- ***Between*** groups design
  - A format where each user experiences a single interface evaluation
- ***Within subjects*** design
  - A format where each user experiences two interface evaluation.

AN EXAMPLE

- Evaluate subparts of the websites of GeorgiaTech and Stanford universities

- GT for GeorgiaTech
- SU for Stanford University

BETWEEN GROUPS FORMAT

- Randomly and evenly divide participants into two groups
- The first group test the GT website
- The second group test the SU website

WITHIN SUBJECTS FORMAT

- Participants test both websites in alternating order
  - GT then SU
  - SU then GA

- Why vary the order?
- What are the advantages and disadvantages of each format?

Strengths of within subjects design

- Power
  - Increase the number of “subjects” relative to a between subjects design
  - More subjects, higher statistical power
  - E.g. if 10 students sign up for the experiment
    - Between groups: 5 subjects in the GA website, and 5 in the SU website
    - Within-subjects: 10 subjects in both websites

STRENGTH CONTINUED

- Reduction in error variance
  - Much of the error variance in a between-group design is the randomness in assigning subjects to groups
    - The two groups may differ with regard to important individual difference factors
  - With within-subject design, the conditions are always exactly equivalent with respect to individual difference variables
  - E.g. in the exercise example, any factor that may affect memory, such as sleep the night before, intelligence, or memory skill, is exactly the same for the two conditions

Weaknesses of within-subjects design

- Carryover effects
  - The participation in one condition may affect performance in other conditions
  - Two basic types of carryover effects
    - Practice: the previous treatment positively effects performance on a later treatment
      - E.g. by dong the task in condition A, the subjects may perform much better the second time in condition B
    - Fatigue: the previous treatment negatively affects performance on a later performance
      - E.g. participants may be tired from two exercises so that they perform worse on the second test

POSSIBLE MITIGATION STRATEGIES

- In a within-subject setting
  - design tasks that are comparable, but not identical, to avoid carry-over effect
  - design tasks that do not take too long to finish, to avoid fatigue problems

- Within-subject studies with 20 subjects (in academic environment), vs., between-group studies(in industry)

COMPARATIVE EVALUATION PROCEDURE

1. Hypothesis development
2. Design of user tasks
3. Measured variables
4. Plan the procedure
5. Recruit the participants
6. Administer the evaluation
7. Results analysis
8. Conclusions

STEP1: HYPOTHESIS DEVELOPMENT

- Hypothesis is a tentative and testable explanation of the relationship between two (or more) events or things; often stated as a prediction that a certain outcome will result from specific conditions.
- Example
  - Hypothesis 1
    - Users spend less time on the GT website downloading the master degree application form
  - Hypothesis 2
    - GT offers a more pleasant experience for master degree applicants

STEP2: DESIGN OF USER TASKS

- Design tasks that help you validate hypotheses
- Consider two hypotheses from previous slide
- Notice the difference

  - Download an application form from the university's website
  - Find the url where you can start submitting the application material

STEP3: MEASURED VARIABLES

- Identify task variables that can be objectively measured
- Time
  - the time from the first mouse button down to the correct target selection
- Error rate
  - the number of wrong item selected vs. the total number of actions required
- Define successful conditions
- http://www.gradadmiss.gatech.edu/apply/apply_now.php
- http://studentaffairs.stanford.edu/gradadmissions/applying/start

STEP4: PLAN THE PROCEDURE

- For each user participant
  - greet her and explain the nature of the evaluation
  - describe the task to her
  - give her the url of two websites to be compared and ***to begin with***
  - ask her to note down the answers for the measured variables
  - congratulate her when she finishes the tasks
  - ask her general questions (you need to prepare a small questionnaire)
  - ask her to comment on the usability (open questions)

STEP5: RECRUIT PARTICIPANTS

- Who?
  - analyze who are the target users of the website/software (interested in applying for master degrees)
  - recruit target users, but vary their experience by beginners vs. intermediates (we prefer beginners because it's usually only one time we visit a website to submit an application)
- Participants are usually recruited, and their time is compensated (cash, course credits, gift vouchers, etc.)
- From where?
  - people you know and their friends or relatives
  - advertise in media platforms (social media, forums)
  - use users from a test centre

STEP6: ADMINISTER THE EVALUATION

1. Supply two evaluators and one user in a given session; one evaluator is the administrator, the other one is the observer (taking notes but doesn't talk)
2. Find a room for the evaluation session; put a "do not disturb" sign on the door
3. Greet the participant
4. Say the orientation script
5. Say the user tasks
6. Record starting time
7. Observer: take notes about interesting and critical events
8. Thank her/him for finishing the tasks
9. Ask post-evaluation questionnaires
10. Ask her/him to provide comments
11. Thank participant, give her gifts, show her out
12. Organize data sheets and notes
13. Summarize thoughts about test
14. Get ready for the next participant

EXAMPLE OF AN ORIENTATION SCRIPT

Taken from previous years

- Hi. My name is Paul. I'll be working with you to test a software. It's a product search tool that allows you to select a Tablet PC. We need your help to find out how user-friendly the system is.
- I'll be asking you to perform some typical tasks with the system. Do your best, but don't be too concerned with results -- the system is being tested, not your performance.

Orientation script cont.

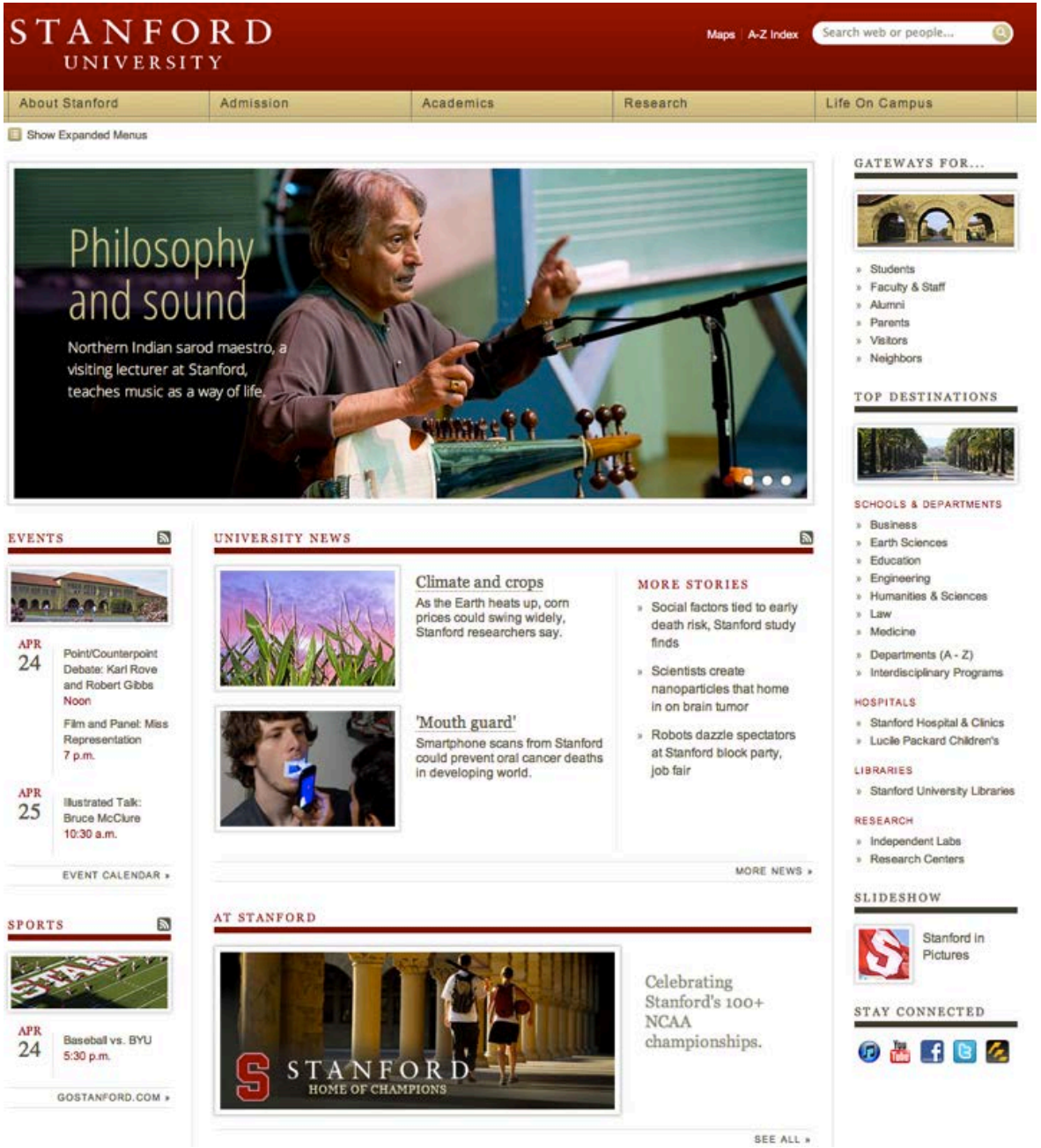
- Please ask questions at any time, but I may only answer them at the end of the session.
- While you are working, my colleague will be taking some notes. We will also be videotaping the session to further analyze the results.
- Sign consent and privacy forms.
- If you feel uncomfortable, you may stop the test at any time.
- Do you have any questions?

STEP7: RESULTS ANALYSIS

- Both results from measured variables and questionnaires need to be analyzed

GEORGIA TECH

STANFORD


STANFORD
UNIVERSITY
Maps | A-Z Index
Search web or people...
About Stanford
Admission
Academics
Research
Life On Campus
Show Expanded Menus
Philosophy and sound
Northern Indian sarod maestro, a visiting lecturer at Stanford, teaches music as a way of life.
GATEWAYS FOR...
Students
Faculty & Staff
Alumni
Parents
Visitors
Neighbors
TOP DESTINATIONS
SCHOOLS & DEPARTMENTS
Business
Earth Sciences
Education
Engineering
Humanities & Sciences
Law
Medicine
Departments (A - Z)
Interdisciplinary Programs
HOSPITALS
Stanford Hospital & Clinics
Lucile Packard Children's
LIBRARIES
Stanford University Libraries
RESEARCH
Independent Labs
Research Centers
SLIDESHOW
Stanford in Pictures
STAY CONNECTED
EVENTS
APR 24
Point/Counterpoint Debate: Karl Rove and Robert Gibbs
Noon
Film and Panel: Miss Representation
7 p.m.
APR 25
Illustrated Talk: Bruce McClure
10:30 a.m.
EVENT CALENDAR »
SPORTS
APR 24
Baseball vs. BYU
5:30 p.m.
GOSTANFORD.COM »
UNIVERSITY NEWS
Climate and crops
As the Earth heats up, corn prices could swing widely, Stanford researchers say.
'Mouth guard'
Smartphone scans from Stanford could prevent oral cancer deaths in developing world.
MORE STORIES
Social factors tied to early death risk, Stanford study finds
Scientists create nanoparticles that home in on brain tumor
Robots dazzle spectators at Stanford block party, job fair
MORE NEWS »
AT STANFORD
STANFORD
HOME OF CHAMPIONS
Celebrating Stanford's 100+ NCAA championships.
SEE ALL »

RESULTS

	Georgia Tech			Stanford		
Participant	Time (sec)	Clicks	Mistakes	Time (sec)	Clicks	Mistakes
1	30	3	0	240	15	3
2	273	7	2	202	11	0
3	300	8	4	180	6	2
4	83	5	0	85	6	0
5	250	21	0	413	36	3
6	112	10	0	101	5	0
Avg.	174	9	1	203	13.17	1.33

p=0.30 (time) p=0.12 (clicks) p=0.36 (mistakes)

RESULTS

	Georgia Tech	Stanford
In which website did you spend more time in finding the application form?	71%	29%
In which website did you use more clicks to accomplish the above task?	43%	57%
Which website has a better structure and organization of its content so that finding the relevant information on each page is easy?	29%	71%
Which website is more pleasant to use?	15%	85%
Which website do you prefer more?	29%	71%

STEP8: CONCLUSIONS

- Users spent on-average more time and clicked more links on the SU website.
- The average errors made are comparable.
- Results are not statistically significant.

- However, users are more satisfied with the SU website: they found the structure and organization better, more pleasant to use, and prefer it.

T-TEST

- The most commonly used method to evaluate the differences in means between two groups
- The t-test can be used even if the sample sizes are very small (e.g., <=10)
- T-test is also known as the student's t-test, since it was published by Gosset in 1908 under the pseudonym "student"

Conclusion

- We presented ***4*** methods
- Critiquing and inspection methods are popular; they can quickly point out usability problems; but require expertise
- Heuristic evaluation offers a guided evaluation with a check-list, but may not cover everything
- Comparative evaluation involves ***a lot of*** users with rigorous procedures (thus requires more effort), but offers objectivity, user perception, and detail
- In industry, designers practice critiquing and heuristic evaluation in early stage of design, and conduct comparative evaluations on critical areas when products are about to roll out

HCI Evaluation should not be an *after-thought*; it should be done during software development

THE END