

Chapter 2

Introducing Evaluation

The aims

- Explain the key concepts used in evaluation.
- Introduce different evaluation methods.
- Show how different methods are used for different purposes at different stages of the design process and in different contexts.
- Show how evaluators mix and modify methods.
- Discuss the practical challenges

Why, what, where and when to evaluate

Iterative design & evaluation is a continuous process that examines:

- Why: to check users' requirements and that users can use the product and they like it.
- What: a conceptual model, early prototypes of a new system and later, more complete prototypes.
- Where: in natural and laboratory settings.
- When: throughout design; finished products can be evaluated to collect information to inform new products.

why you need to evaluate

Iterative design, with its repeating cycle of design and testing, is the only validated methodology in existence that will consistently produce successful results. If you don't have user-testing as an integral part of your design process you are going to throw buckets of money down the drain.

why you need to evaluate

There is a need to invest in evaluation because:

- users expect usable system as well as pleasing and enjoyable experience.
- From a business and marketing perspective, only well-designed products sell.
- Designers need to focus on real problems and the needs of different user groups and not what they like or dislike.

Activity

Identify one adult and one teenager prepared to talk about with you about their facebook usage.

Ask them questions such as:

- How often do they post photo?
- What kind of photos?
- How many friends have you got?
- What applications do you have?
- Have you befriended anyone?

Objectives of User Interface Evaluation

- Key objective of both UI design and evaluation:

Minimize malfunctions

- Key reason for focusing on evaluation:
 - Without it, the designer would be working “blindfold”
 - Designers wouldn’t really know whether they are solving customer’s problems in the most productive way

Objectives of User Interface Evaluation

Questions answered by various evaluation techniques:

1. What is the user's real task?
 - Prevent later malfunctions
 - by doing evaluation as part of requirements analysis
 - Present and work with a UI
 - to help formulate the requirements
 - Inappropriate tasks/requirements are a major source of malfunctions
2. What problems do or might users experience with the UI?
 - Directly find malfunctions
3. Which of several alternative UI's is better?
 - Pick the version that leads to fewer malfunctions
4. Has the UI met usability targets?
 - Ensure that malfunction counts are sufficiently low
5. Does the UI conform to standards?
 - Leverage of collective wisdom to reduce malfunctions
6. Decide how to deal with the ethical issues
 - Consent form

Objectives of User Interface Evaluation

- But, in order for evaluation to give feedback to designers...
- ...we must understand why a malfunction occurs
- Malfunction analysis:
 - Determine why a malfunction occurs
 - Determine how to eliminate malfunctions

Malfunction Analysis

- A disciplined approach to analyzing malfunctions
 - Provides feedback into the redesign process
 1. Play protocol, searching for malfunctions
 2. Answer four distinct questions:
 - Q1. How is the malfunction manifested?
 - What do you notice and who noticed it?
 - Q2. At what stage in the interaction is it occurring?
 - Goal forming, action decision, action execution, interpretation of results
 - Q3. At what level of the user interface is it occurring?
 - Physical element level to task level
 - Q4. Why is it occurring?
 - What is its root cause
 3. List and prioritize possible cures

Q1. How is the malfunction manifested?

- a) Malfunctions detected by the system (easiest to detect)
 - omission of an argument
 - incorrect date format
- Cure:
 - Better prompts, consistency, visible examples, more forgiving of alternatives
- b) Malfunctions detected by the user during operation
 - taking wrong path in menu hierarchy
 - not finding required help
 - not being able to perform a certain action
 - not being able to tell which state system is in
- Cure:
 - Improve functionality, feedback, clarity, simplicity

Q1. How is the malfunction manifested? (cont'd)

- c) Malfunctions undetected (until later)
 - output produced is wrong due to wrong inputs
 - unnecessary work performed
- Cure:
 - Improve feedback indicating consequences of input; simplify
- d) Inefficiencies
 - excessive response time
 - excessive think time
 - unnecessarily long command sequences
 - unnecessary repetitions
 - complex operations that require use of reference
- Cure:
 - Simplify, speed system up

Q2. What Stage in the Interaction the Malfunction Occur?

- a) When the user decides on next goal (Forms an intent to do inappropriate thing)
 - decides to empty a field because user thinks it is unimportant (when it is important)
 - decides to charge default exchange rate (when should obtain current exchange rate)
- Cure:
 - Lead user through task better; better feedback; better training
- b) When the user specifies the action (Action does not match the goal)
 - deletes the record instead of emptying a field
- Cure:
 - Improve clarity, feedback, prompts, conceptual model

Q2. What Stage in the Interaction the Malfunction Occur? (cont'd)

- c) When the system executes the action
 - Defects in functionality
- Cure:
 - Fix functionality in normal way
- d) When the user interprets the resulting system state
 - thinks bank account has been debited when it has not
 - thinks system has 'hung' when it has not
 - thinks some data must be entered when it is the default
 - cannot understand resulting error message
- Cure:
 - Better feedback, better conceptual model

Q3. At Which Level Does the Malfunction Occur?

- a) Task level (Task and goals not supported)
 - What the user wants to do cannot be done by the system
 - Functionality is not provided
- Cure:
 - Add functionality
- b) Conceptual level (User has wrong mental model; does not understand intended conceptual model)
 - thinks that money is being deducted from bank account when it is being charged to a credit card
 - thinks that dragging a file to the desktop means they are no longer on the disk
 - thinks that dragging a disk to the trash can icon deletes disk contents
- Cure:
 - make conceptual model clearer; improve metaphors

Q3. At Which Level Does the Malfunction Occur? (cont'd)

- c) Interaction style level (system wide problem)
 - does not know how to pull down a menu
 - scrolls a page instead of a line
 - goes to next screen instead of scrolling
 - retypes command after an error instead of editing it
- Cure:
 - make operation of the interface more intuitive and consistent
- d) Interaction element level (specific detail inappropriate)
 - selects wrong button because label is misinterpreted
 - specifies invalid command syntax
 - specifies wrong code for option
- Cure:
 - More attention to details of the interface, simplification

Q3. At Which Level Does the Malfunction Occur? (cont'd)

- e) Physical element level (Physical execution incorrect)
 - presses wrong key accidentally
 - clicks on wrong pixel in image
 - types ahead when system is computing; keystrokes later applied to wrong action
- Cure:
 - Defenses to protect user from consequences; better hardware design; fix bugs in code

Q4. Why Does the Malfunction Occur?

- a) Lack of (on the part of the user):
 - Motivation:
 - Poor job satisfaction
 - Attention:
 - User is pre-occupied with other things.
 - Input information processing:
 - No feedback provided to tell user what is going on
 - or cues provided by the system are not recognized
 - or cues are misinterpreted
 - Cures: Clearer, more consistent feedback
 - Discrimination:
 - user is unable to tell certain things apart
 - e.g. red/green colour discrimination
 - e.g. two icons that are similar
 - Cures: Improved expression of information

Q4. Why Does the Malfunction Occur? (cont'd)

- Physical coordination:
 - e.g. wrong item selected because of difficulty positioning cursor with mouse.
- Cures: Alternate interaction mechanisms, better feedback
- Recall:
 - User did not remember command , syntax etc.
- Cures: Better mnemonics, online help, quick lookup mechanisms, command completion
- Knowledge / lack of learning:
 - User does not have business or software knowledge to make right choice.

Q4. Why Does the Malfunction Occur? (cont'd)

- b) Learning difficulties that cause malfunctions:
 - Learning is difficult
 - users get frustrated
 - learning takes time; can be hard to apply
 - Learners make ad-hoc interpretations
 - they may not recognize their problem
 - they may falsely think they have a problem
 - Learners generalize from what they know
 - they assume computers work like manual methods
 - they assume consistency
 - Learners have trouble following directions
 - they often ignore them even if they see them
 - they do not easily understand them

Q4. Why Does the Malfunction Occur? (cont'd)

- b) Learning difficulties that cause malfunctions:
 - Problems and features interact
 - they do not see that one problem can cause another
 - Prerequisites and side-effects confuse learners
 - Help facilities do not always help
 - they do not know what to ask for
 - too much detail is often provided
 - Other causes of malfunctions:
 - Excessive resource demands
 - External events (e.g. noise)
 - Misleading or inadequate training
 - Unrealistic task definitions
 - Intrinsic human variability

What to evaluate

- Ranges from Low-tech prototypes to complete systems, from aesthetic design to safety features.
- Examples of what evaluate:
 - new web browser
 - Ambient display
 - Computerized systems for controlling traffic lights, to see if it results in few accidents, etc.
 - Software company may want to assess market reaction to its homepage design.
 - Why would you evaluate a toy, or a digital music player?

Activity

- What aspects would you want to evaluate for the following systems:
 1. A personal music player (e.g. iPod)?
 2. A website for selling clothes?

Where to evaluate

- Laboratories: provides the control necessary to investigate if requirements are met. (e.g. web accessibility, design choices for layout/size of keys in a cell phone)
- Natural Settings or wild studies. (e.g. children enjoy playing with a new toy, social networking can be evaluated at home).
- Living Laboratories: Between regular labs and natural settings. (e.g. home turned into an environment where activities can be recorded, measured and controlled).

Activity

- A company is developing a new car seat to monitor if a person starts to fall asleep when driving and to provide a wake-up call using haptic feedback or other means.
1. Where would you evaluate it?

When to evaluate

At what stage in product life cycle evaluation takes place depends on type of product.

Formative evaluations: When evaluations are done during design to check if product continues to meet user's need.

1. if product being developed is a brand new concept.
2. considerable time is invested in market research to establish user requirements.
3. requirements are used to create initial sketches, storyboard, series of screens, prototype of design ideas.

- **Summative evaluations:** when evaluations are done to assess the success of a finished product.

1. evaluation will ascertain what needs improving.
2. new features need to be added or
3. existing features need to be improved.

Types of evaluation

- Controlled settings involving users: user's activities are controlled to test hypothesis and measure and observe certain behaviors. (main methods are usability testing & experiments in laboratories and living labs).
- Natural settings involving users: there is little or no control of user's activities to see how the product is used in the real world (main method is field studies in public places or online communities).
- Any settings not involving users, eg consultants critique; to predict, analyze & model aspects of the interface analytics to identify the most usability problems. (main methods are inspections, heuristics, walkthroughs, models, and analytics).

Pros and cons of each type

- Lab-based studies are good at revealing usability problems but poor at capturing context of use.
- Field studies are good at demonstrating how people use technologies in their intended setting, but are expensive and difficult to conduct.
- Modeling and predicting approaches are cheap and quick to perform, but can miss unpredictable usability problems and subtle aspects of the user experience.

Activity

Imagine you have designed a website for teenagers to share music, gossip, and photos. You have prototyped your first design and implemented the core functionality.

How would you find out if it would appeal to them and how they use it to carry out the tasks?
what type of evaluation?

- To decide which approach/type to use depends on how much control is needed in order to find out how an interface/device is used.

Controlled settings involving users

Enable evaluators to control what users do, when they do it, and for how long. (e.g. used to evaluate PC software apps where participants can be seated in front of them)

- Usability testing:

1. Collecting data using a combination of methods such as observation, interviews, questionnaires in a controlled setting.
2. Done in laboratories but increasingly is done remotely or in natural settings
3. Primary goal is to determine the usability of the interface by the intended user population to carry out tasks for which it was designed.
4. Recording is done on video or by logging software.
5. Supplemented by observation at product sites to collect evidence about how product is used.

Controlled settings involving users

- Experiments: (e.g . Comparing which is best way for users to enter text when using virtual keyboard vs physical keyboard)
 1. Done in research laboratories.
 2. Most controlled settings where researchers try to remove any extraneous variables that may interfere with the participants performance.
 3. Primary goal is to test hypothesis. (e.g. whether one is better than the other in terms of speed of typing, number of errors, etc.
 4. Data is collected and analyzed

Usability lab



http://iat.ubalt.edu/usability_lab/

Living labs

- People's use of technology in their everyday lives can be evaluated in living labs.
- Such evaluations are too difficult to do in a usability lab.
- Eg the Aware Home was embedded with a complex network of sensors and audio/video recording devices that recorded their movements and their use of technology.

Natural settings involving users

- Evaluate people in their natural settings. Used primarily to:
 1. Help identify opportunities for new technology
 2. Establish requirements for new design
 3. Facilitate the introduction of technology
- Data is collected by observation, interviews and logging in the form of diary notes by participants, video/audio recording or notes by researchers.
- Goal is to be unobtrusive and not to affect what people do during the evaluation.
- Trend towards wild studies (homes, public places, outdoors)
- Researchers inevitably have to give up control of what is being evaluated.
- Need to recruit people who are willing to use a device to be evaluated for few weeks/months.

Activity

- What is the downside of handing over control to participants in a Natural settings evaluations?
- Field studies can also be virtual where observations take place in multiuser games, chat rooms, and so on.
- Examines kinds of social processes such as cooperation, confrontation, collaboration.

Any settings not involving users

Evaluations are conducted in settings where the researcher has to imagine or model how an interface is likely to be used.

Methods include

Inspections : Commonly employed to predict user's behavior and to identify usability problems based on knowledge of usability, user's behavior, context in which system will be used and kinds of activities users undertake. Done by experts.

1. Heuristic Evaluation: Experts, guided by a set of usability principles (known as heuristics), evaluate whether UI elements conform to tried and tested principles
2. Cognitive walkthroughs: Walking through a task with the product and noting problematic usability features

Any settings not involving users

Analytics: Technique for logging data either at a customer's site or remotely. It is a method for evaluating traffic through a system. (e.g. Web analytics is the measurement, collection, analysis and reporting internet data in order to understand and optimize web usage).

Models: Technique used primarily for comparing the efficacy of different interfaces for the same application.

(e.g. optimal arrangement and location of features).

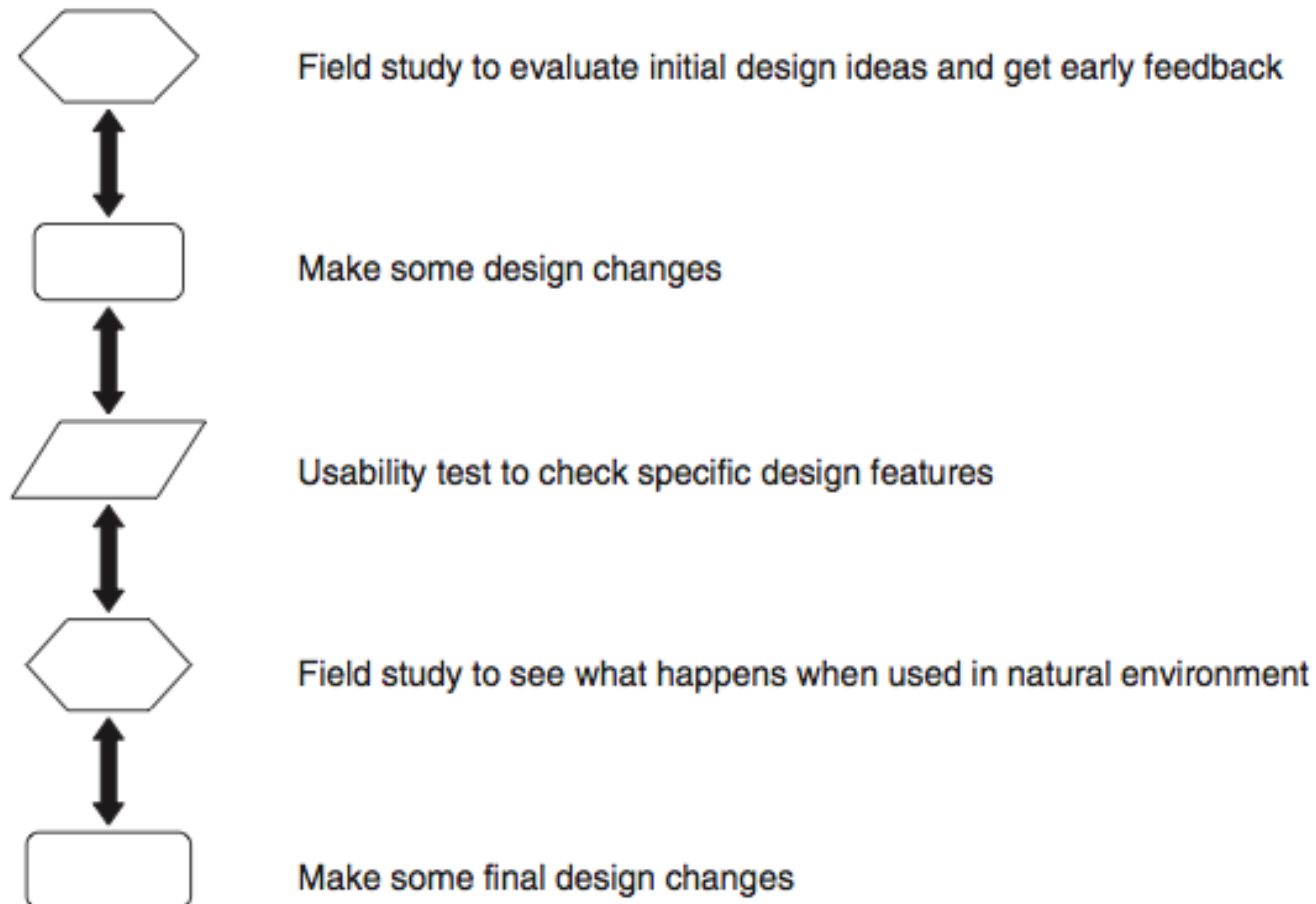
Combination of methods

The three broad categories provide a general framework to guide the selection of evaluation methods

Often Combinations of methods are used across these categories to obtain a richer understanding.

For example, usability testing in labs can be combined with observations in natural settings to identify usability problems. (testing a cell phone)

Usability testing & field studies can compliment



Evaluation case studies

- Hotel Reservations
- Experiment to investigate a computer game
- In the wild field study of skiers
- Crowdsourcing

A Preliminary Case Study: Hotel Reservations

- UI Evaluation performed for Forte Travelodge Performed in a special usability lab
- Aims:
 - Identify and eliminate malfunctions
 - Hence make system easier to use
 - Avoid business difficulties caused by these malfunctions
 - Develop improved training material and documentation
 - Avert potential malfunctions by teaching users how to avoid them
- Setup of IBM usability lab:
 - Resembles TV studio
 - Microphones and video equipment
 - One way mirror
 - Technicians, observers sit on one side
 - Users sit on other side in realistic environment
 - User environment resembles reception desk
 - Non-threatening

A Preliminary Case Study: Hotel Reservations

- Aspects of system to be evaluated:
 - How quickly can a booking be made?
 - (while operator is on telephone)
 - Is each screen productive to use?
 - Are help and error messages effective?
 - Can non-computer-literate operators use the system?
 - Is complexity minimized?
 - Is training and documentation effective?


A Preliminary Case Study: Hotel Reservations

- Procedure:

- 15 common task scenarios developed:
 - Among others: basic registration, cancellation, request for specific room, extension of existing stay etc.
- Four days of testing with multiple users performing various sets of tasks
 - Users were told evaluation is of system, not them
 - All actions were recorded
 - Debriefing sessions held
- Videos then analyzed for malfunctions
 - 62 identified
 - Priorities:
 - Navigation speed needs improvement
 - Screen titles and formats need tuning
 - Hard to refer to documentation
 - Physical difficulties with telephone headsets and furniture

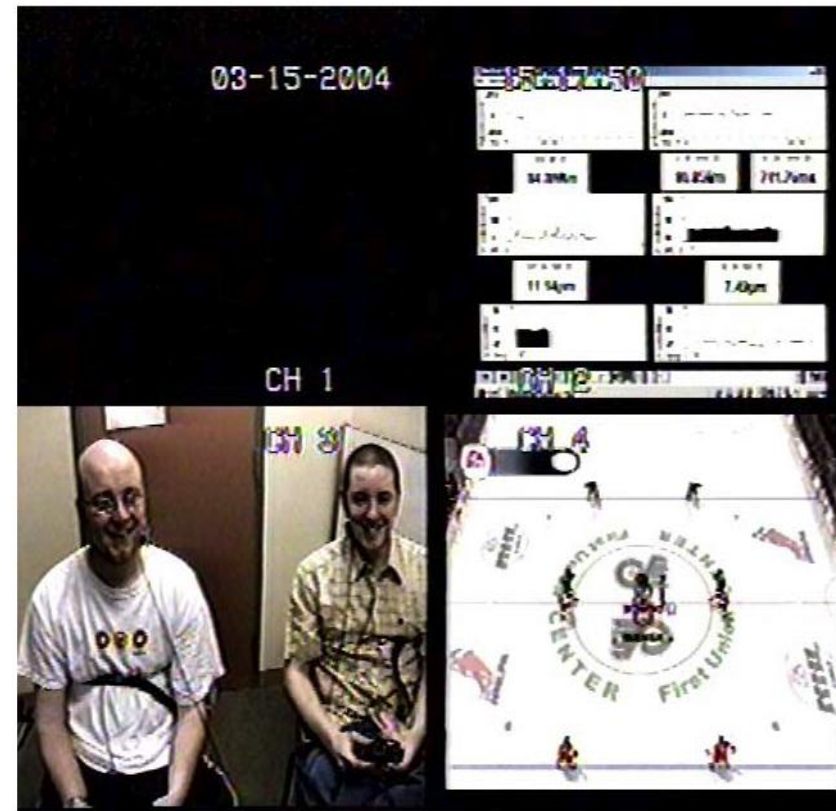
A Preliminary Case Study: Hotel Reservations

- Results:

- Higher productivity of booking staff
 - tasks completed more quickly
 - guest requirements better met
- Training costs kept low
- Morale kept high
- More customers booked by phone
 - 14500  27000 per week

Challenge & engagement in a collaborative immersive game

- Physiological measures were used.
- Players were more engaged when playing against another person than when playing against a computer.
- What precautionary measures did the evaluators take?



What does this data tell you?

high values indicate more variation

	Playing against computer		Playing against friend	
	Mean	St. Dev.	Mean	St. Dev.
Boring	2.3	0.949	1.7	0.949
Challenging	3.6	1.08	3.9	0.994
Easy	2.7	0.823	2.5	0.850
Engaging	3.8	0.422	4.3	0.675
Exciting	3.5	0.527	4.1	0.568
Frustrating	2.8	1.14	2.5	0.850
Fun	3.9	0.738	4.6	0.699

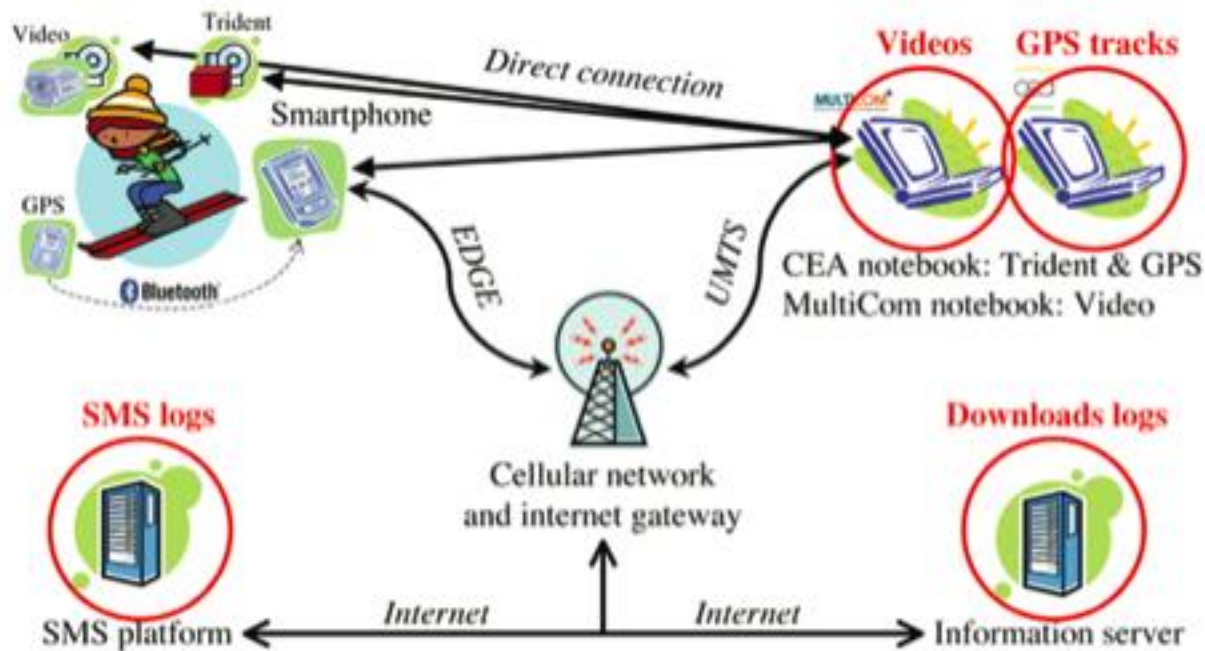
Source: Mandryk and Inkpen (2004).

Why study skiers in the wild ?



Jambon et al. (2009) User experience in the wild. In: Proceedings of CHI '09, ACM Press, New York, p. 4070-4071.

e-skiing system components



Jambon et al. (2009) *User experience in the wild*. In: *Proceedings of CHI '09*, ACM Press, New York, p. 4072.

Crowdsourcing-when might you use it?

amazon mechanical turk
Artificial Intelligence

Your Account

HITs

Qualifications

Already have an account?
Sign in as a [Worker](#) | [Requester](#)

[Introduction](#) | [Dashboard](#) | [Status](#) | [Account Settings](#)

Mechanical Turk is a marketplace for work.

We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

161,325 HITs available. [View them now.](#)

Make Money by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. [Find HITs now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



or [learn more about being a Worker](#)

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. [Register Now](#)

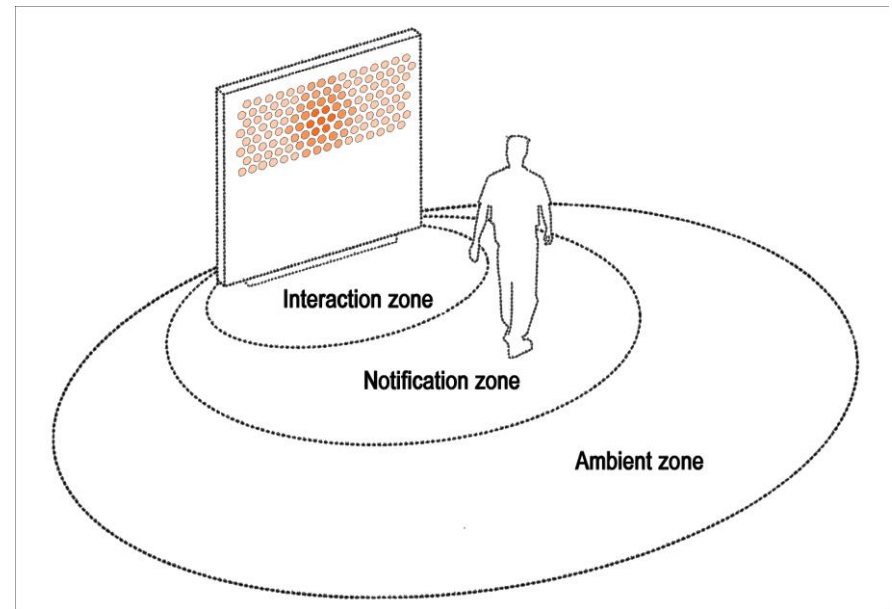
As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results



Evaluating an ambient system

- The Hello Wall is a new kind of system that is designed to explore how people react to its presence.
- What are the challenges of evaluating systems like this?



Evaluation methods

Method	Controlled settings	Natural settings	Without users
Observing	X	X	
Asking users	X	X	
Asking experts		X	X
Testing	X		
Modeling			X

The language of evaluation

Analytics

Analytical
evaluation

Controlled
experiment

Expert review or crit

Field study

Formative
evaluation

Heuristic evaluation

In the wild
evaluation

Living laboratory

Predictive evaluation

Summative
evaluation

Usability laboratory

User studies

Usability testing

Users or participants

Key points

- Evaluation & design are closely integrated in user-centered design.
- Some of the same techniques are used in evaluation as for establishing requirements but they are used differently (e.g. observation interviews & questionnaires).
- Three types of evaluation: laboratory based with users, in the field with users, studies that do not involve users
- The main methods are: observing, asking users, asking experts, user testing, inspection, and modeling users' task performance, analytics.
- Dealing with constraints is an important skill for evaluators to develop.

THANK YOU

ΕΥΧΑΡΙΣΤΩ

谢谢

Merci

ขอบคุณ

Vielen
Dank

DMnvwd

Gracias

شكراً

Grazie

Hvala

Bedankt

Dankie

ڳو رايٻھ ماٿيھ اڱايٻھ

Obrigado!

Dikey

Köszönettel

ありがとう

WAD MAHAD

SAN TAHAY



Dziękuję

СПАСИБО

감사합니다

Tessekkürler

متشكرم

GADDA GUEY

Asante Urakoze