Chapter 12

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.

Bruce Tognazzini tells you 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."

See AskTog.com for topical discussions about design and evaluation.

Types of evaluation

- Controlled settings involving users, eg usability testing & experiments in laboratories and living labs.
- Natural settings involving users, e.g. field studies to see how the product is used in the real world.
- Any settings not involving users, e.g. consultants critique; to predict, analyze & model aspects of the interface analytics.

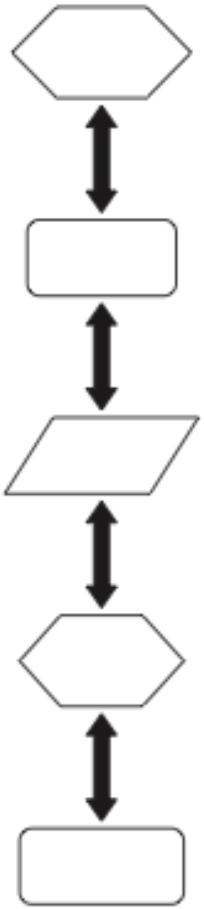
Usability lab

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.
- e.g. the Aware Home was embedded with a complex network of sensors and audio/video recording devices (Abowd et al., 2000).

Usability testing & field studies can compliment



Field study to evaluate initial design ideas and get early feedback

Make some design changes

Usability test to check specific design features

Field study to see what happens when used in natural environment

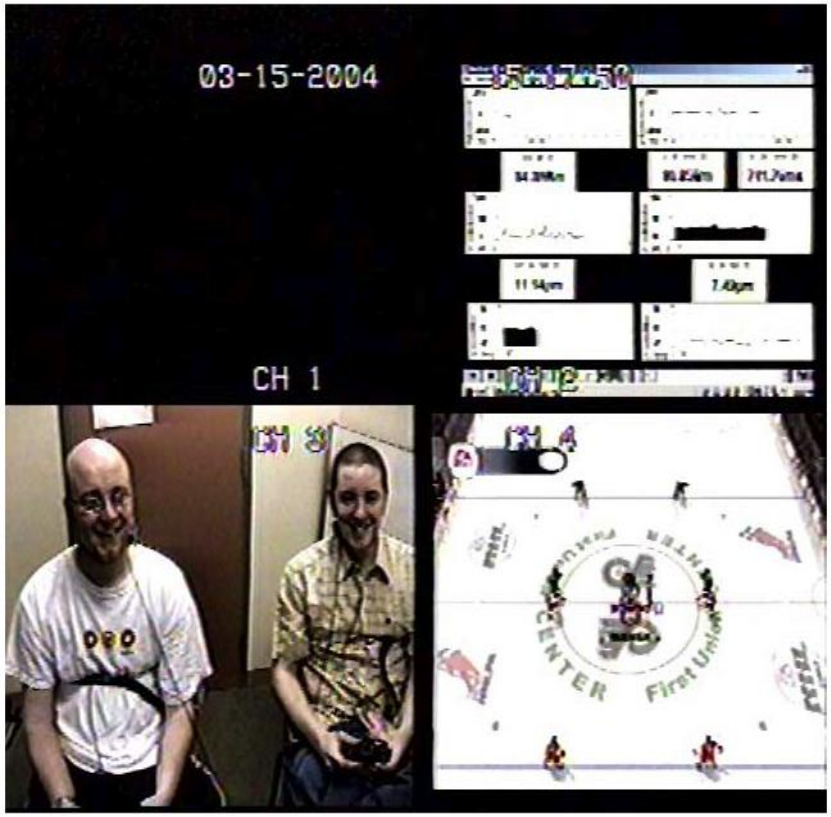
Make some final design changes

Evaluation case studies

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

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

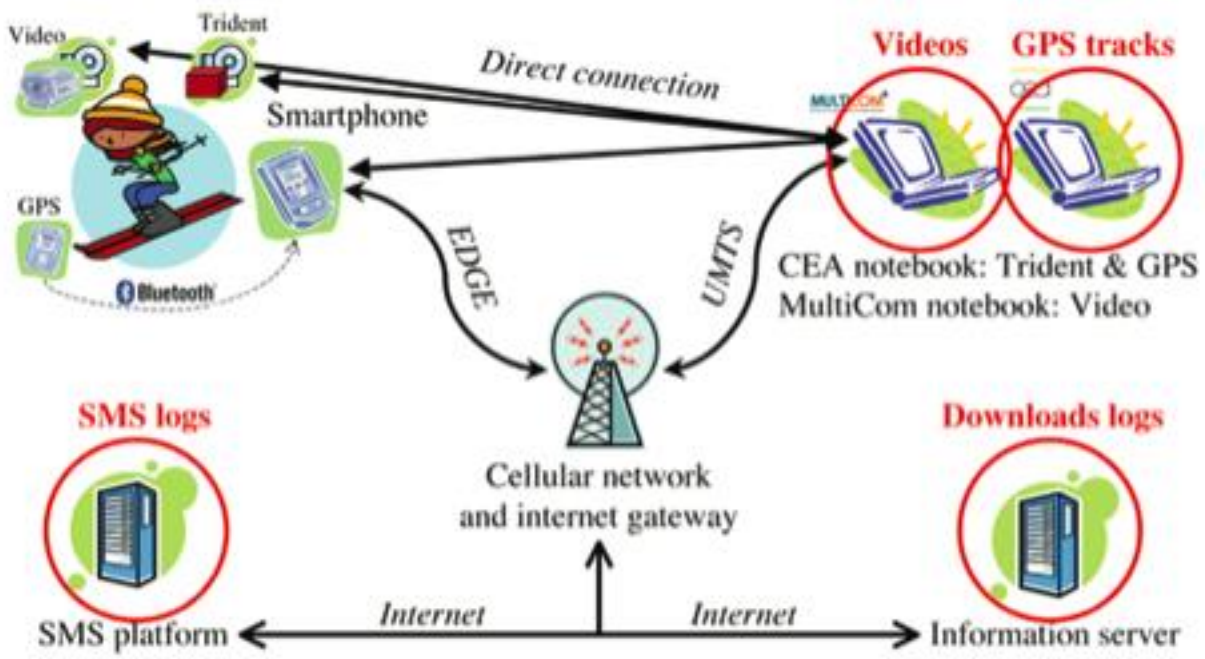
	Playing against computer		Playing against friend	
	Mean	St. Dev.	Mean	St. Dev.
Boring	2.3	0.949	1.7	0.949
Challenging	3.6	1.08	3.9	0.994
Easy	2.7	0.823	2.5	0.850
Engaging	3.8	0.422	4.3	0.675
Exciting	3.5	0.527	4.1	0.568
Frustrating	2.8	1.14	2.5	0.850
Fun	3.9	0.738	4.6	0.699

Source: Mandryk and Inkpen (2004).

Why study skiers in the wild ?

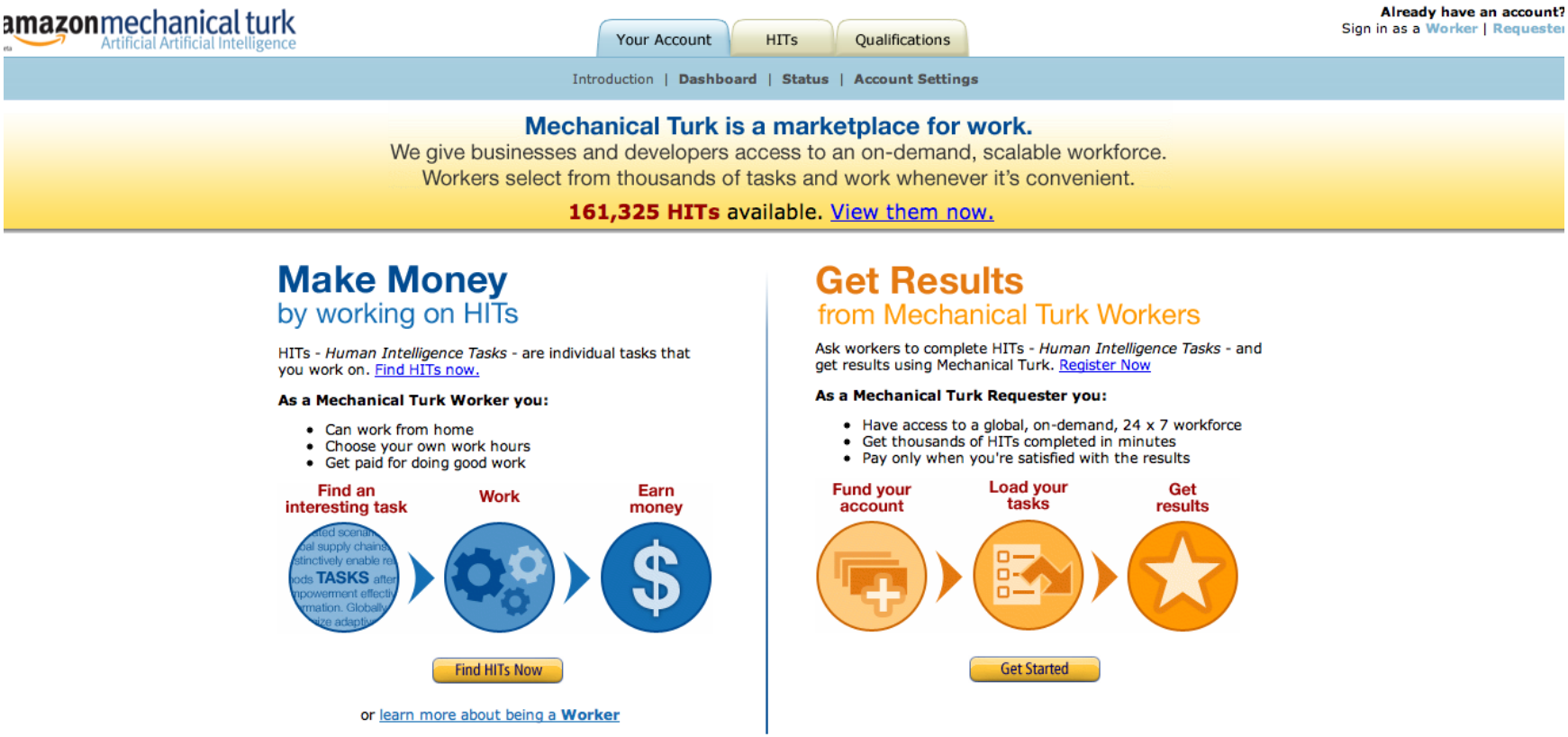
Jambon et al. (2009) User experience in the wild. In: Proceedings of CHI '09, ACM Press, New York, p. 4070-4071.

e-skiing system components



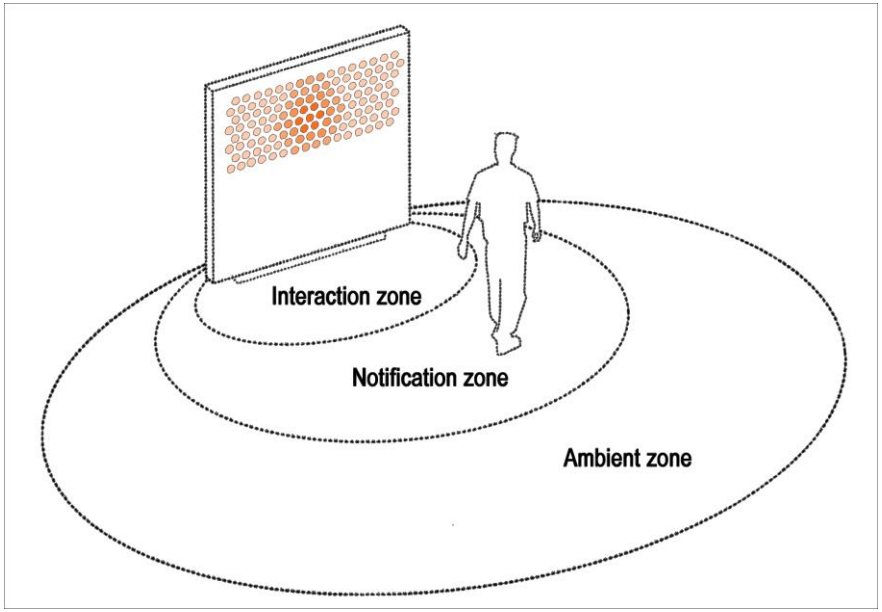
Jambon et al. (2009) User experience in the wild. In: Proceedings of CHI '09, ACM Press, New York, p. 4072.

Crowdsourcing-when might you use it?



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.

Chapter 13

An evaluation framework

The aims are:

- Introduce and explain the DECIDE framework.
- Discuss the conceptual, practical, and ethical issues involved in evaluation.

DECIDE: a framework to guide evaluation

- Determine the *goals*.
- Explore the *questions*.
- Choose the evaluation *methods*.
- Identify the *practical issues*.
- Decide how to deal with the *ethical issues.*
- Evaluate, analyze, interpret and present the *data*.

Determine the goals

- What are the high-level goals of the evaluation?
- Who wants it and why?
- The goals influence the methods used for the study.
- Goals vary and could be to:
  - identify the best metaphor for the design
  - check that user requirements are met
  - check for consistency
  - investigate how technology affects working practices
  - improve the usability of an existing product

Explore the questions

- Questions help to guide the evaluation.
- The goal of finding out why some customers prefer to purchase paper airline tickets rather than e-tickets can be broken down into sub-questions:
  - What are customers' attitudes to e-tickets?
  - Are they concerned about security?
  - Is the interface for obtaining them poor?
- What questions might you ask about the design of a cell phone?

Choose the evaluation approach & methods

- The evaluation method influences how data is collected, analyzed and presented.
- E.g. field studies typically:
  - Involve observation and interviews.
  - Involve users in natural settings.
  - Do not involve controlled tests.
  - Produce qualitative data.

Identify practical issues

For example, how to:

- Select users
- Find evaluators
- Select equipment
- Stay on budget
- Stay on schedule

Decide about ethical issues

- Develop an informed consent form

- Participants have a right to:
  - Know the goals of the study;
  - Know what will happen to the findings;
  - Privacy of personal information;
  - Leave when they wish;
  - Be treated politely.

Evaluate, interpret & present data

- Methods used influence how data is evaluated, interpreted and presented.
- The following need to be considered:
  - Reliability: can the study be replicated?
  - Validity: is it measuring what you expected?
  - Biases: is the process creating biases?
  - Scope: can the findings be generalized?
  - Ecological validity: is the environment influencing the findings? i.e. Hawthorn effect.

Key points

- Many issues to consider before conducting an evaluation study.
- These include: goals of the study; involvment or not of users; the methods to use; practical & ethical issues; how data will be collected, analyzed & presented.
- The DECIDE framework provides a useful checklist for planning an evaluation study.

Chapter 14
Evaluation studies: From controlled to natural settings

The aims:

- Explain how to do usability testing
- Outline the basics of experimental design
- Describe how to do field studies

Usability testing

- Involves recording performance of typical users doing typical tasks.
- Controlled settings.
- Users are observed and timed.
- Data is recorded on video & key presses are logged.
- The data is used to calculate performance times, and to identify & explain errors.
- User satisfaction is evaluated using questionnaires & interviews.
- Field observations may be used to provide contextual understanding.

Experiments & usability testing

- Experiments test hypotheses to discover new knowledge by investigating the relationship between two or more things – i.e., variables.
- Usability testing is applied experimentation.
- Developers check that the system is usable by the intended user population for their tasks.
- Experiments may also be done in usability testing.

Usability testing & research

Usability testing

- Improve products
- Few participants
- Results inform design
- Usually not completely replicable
- Conditions controlled as much as possible
- Procedure planned
- Results reported to developers

Experiments for research

- Discover knowledge
- Many participants
- Results validated statistically
- Must be replicable
- Strongly controlled conditions
- Experimental design
- Scientific report to scientific community

Usability testing

- Goals & questions focus on how well users perform tasks with the product.
- Comparison of products or prototypes common.
- Focus is on time to complete task & number & type of errors.
- Data collected by video & interaction logging.
- Testing is central.
- User satisfaction questionnaires & interviews provide data about users' opinions.

Usability lab with observers watching a user & assistant

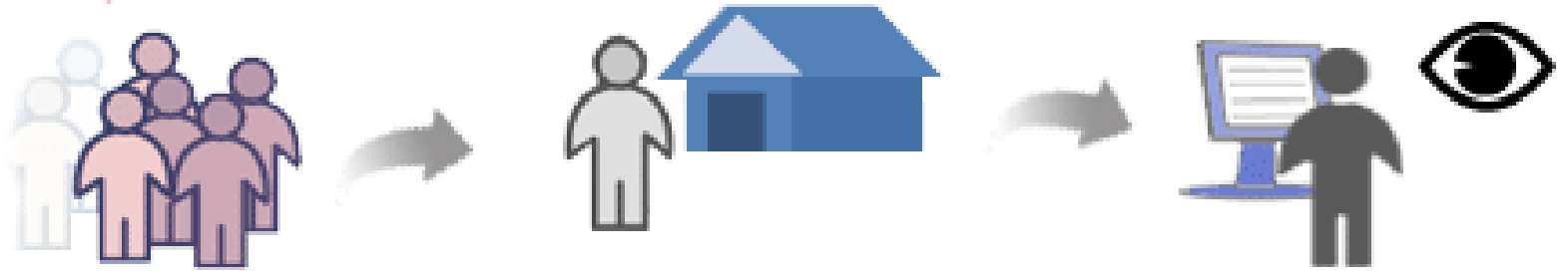
Portable equipment for use in the field

Remote Usability Testing

A selected group of panelists are invited to participate

...They are asked to evaluate the web from their natural context, using Internet Explorer

...A robot (UZ Bar) guides the users and monitors their behavior



The users are asked to complete certain tasks and answer questions

The UZ Platform gatheres and saves the data in real-time

Servidores UserZoom

The data is analysed and a final report is prepared

Mobile head-mounted eye tracker

Picture courtesy of SensoMotoric Instruments (SMI), copyright 2010

Testing conditions

- Usability lab or other controlled space.
- Emphasis on:
  - selecting representative users;
  - developing representative tasks.
- 5-10 users typically selected.
- Tasks usually last no more than 30 minutes.
- The test conditions should be the same for every participant.
- Informed consent form explains procedures and deals with ethical issues.

Some type of data

- Time to complete a task.
- Time to complete a task after a specified. time away from the product.
- Number and type of errors per task.
- Number of errors per unit of time.
- Number of navigations to online help or manuals.
- Number of users making a particular error.
- Number of users completing task successfully.

Usability engineering orientation

- Aim is improvement with each version.
- Current level of performance.
- Minimum acceptable level of performance.
- Target level of performance.

How many participants is enough for user testing?

- The number is a practical issue.
- Depends on:
  - schedule for testing;
  - availability of participants;
  - cost of running tests.
- Typically 5-10 participants.
- Some experts argue that testing should continue until no new insights are gained.

Experiments

- Predict the relationship between two or more variables.
- Independent variable is manipulated by the researcher.
- Dependent variable depends on the independent variable.
- Typical experimental designs have one or two independent variable.
- Validated statistically & replicable.

Experimental designs

- Different participants - single group of participants is allocated randomly to the experimental conditions.
- Same participants - all participants appear in both conditions.
- Matched participants - participants are matched in pairs, e.g., based on expertise, gender, etc.

Different, same, matched participant design

Design	Advantages	Disadvantages
Different	No order effects	Many subjects & individual differences a problem
Same	Few individuals, no individual differences	Counter-balancing needed because of ordering effects
Matched	Same as different participants but individual differences reduced	Cannot be sure of perfect matching on all differences

Field studies

- Field studies are done in natural settings.
- “in the wild” is a term for prototypes being used freely in natural settings.
- Aim to understand what users do naturally and how technology impacts them.
- Field studies are used in product design to:
  - identify opportunities for new technology;
  - determine design requirements;
  - decide how best to introduce new technology;
  - evaluate technology in use.

Data collection & analysis

- Observation & interviews
  - Notes, pictures, recordings
  - Video
  - Logging
- Analyzes
  - Categorized
  - Categories can be provided by theory
    - Grounded theory
    - Activity theory

Data presentation

- The aim is to show how the products are being appropriated and integrated into their surroundings.
- Typical presentation forms include: vignettes, excerpts, critical incidents, patterns, and narratives.

UbiFit Garden: An in the wild study

Key points

- Usability testing is done in controlled conditions.
- Usability testing is an adapted form of experimentation.
- Experiments aim to test hypotheses by manipulating certain variables while keeping others constant.
- The experimenter controls the independent variable(s) but not the dependent variable(s).
- There are three types of experimental design: different-participants, same- participants, & matched participants.
- Field studies are done in natural environments.
- "In the wild" is a recent term for studies in which a prototype is freely used in a natural setting.
- Typically observation and interviews are used to collect field studies data.
- Data is usually presented as anecdotes, excerpts, critical incidents, patterns and narratives.

Chapter 15

Analytical evaluation

Aims:

- Describe the key concepts associated with inspection methods.
- Explain how to do heuristic evaluation and walkthroughs.
- Explain the role of analytics in evaluation.
- Describe how to perform two types of predictive methods, GOMS and Fitts' Law.

Inspections

- Several kinds.
- Experts use their knowledge of users & technology to review software usability.
- Expert critiques (crits) can be formal or informal reports.
- Heuristic evaluation is a review guided by a set of heuristics.
- Walkthroughs involve stepping through a pre-planned scenario noting potential problems.

Heuristic evaluation

- Developed by Jacob Nielsen in the early 1990s.
- Based on heuristics distilled from an empirical analysis of 249 usability problems.
- These heuristics have been revised for current technology.
- Heuristics being developed for mobile devices, wearables, virtual worlds, etc.
- Design guidelines form a basis for developing heuristics.

Nielsen's original heuristics - 1

- **Visibility of system status.**

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

- **Match between system and 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.

- **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.

- **Consistency and standards.**

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

Nielsen's original heuristics - 2

- **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.

- **Recognition rather than recall.**

  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.

- **Flexibility and efficiency of use.**

  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.

- **Aesthetic and minimalist design.**

  Dialogues 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.

Nielsen's original heuristics - 3

- **Help users recognize, diagnose, recover from errors.**

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

- **Help and documentation.**

  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.

Nielsen, J. (1994a). Enhancing the explanatory power of usability heuristics. Proc. ACM CHI'94 Conf. (Boston, MA, April 24-28), 152-158.

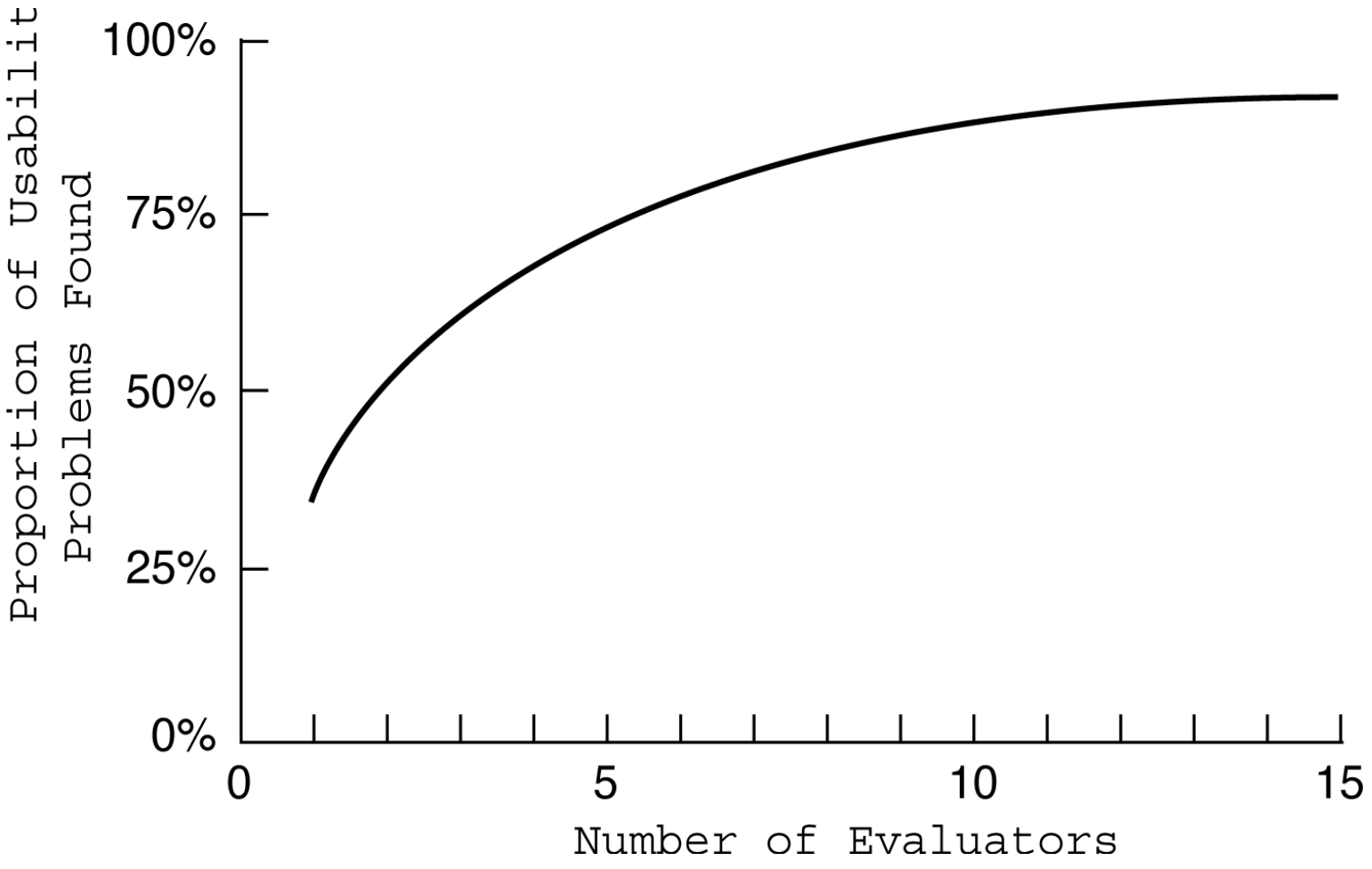
Nielsen, J. (1994b). Heuristic evaluation. In Nielsen, J., and Mack, R.L. (Eds.), Usability Inspection Methods, John Wiley & Sons, New York, NY.

Discount evaluation

- Heuristic evaluation is referred to as discount evaluation when 5 evaluators are used.

- Empirical evidence suggests that on average 5 evaluators identify 75-80% of usability problems.

No. of evaluators & problems



3 stages for doing heuristic evaluation

- Briefing session to tell experts what to do.
- Evaluation period of 1-2 hours in which:
  - Each expert works separately;
  - Take one pass to get a feel for the product;
  - Take a second pass to focus on specific features.
- Debriefing session in which experts work together to prioritize problems.

Advantages and problems

- Few ethical & practical issues to consider because users not involved.
- Can be difficult & expensive to find experts.
- Best experts have knowledge of application domain & users.
- Biggest problems:
  - Important problems may get missed;
  - Many trivial problems are often identified;
  - Experts have biases.

Heuristics for websites focus on key criteria (Budd, 2007)

- Clarity
- Minimize unnecessary complexity & cognitive load
- Provide users with context
- Promote positive & pleasurable user experience

[Budd 2007]
http://www.andybudd.com/archives/2007/01/heuristics_for_modern_web_application_development/

Cognitive walkthroughs

- Focus on ease of learning.
- Designer presents an aspect of the design & usage scenarios.
- Expert is told the assumptions about user population, context of use, task details.
- One or more experts walk through the design prototype with the scenario.
- Experts are guided by 3 questions.

The 3 questions

- Will the correct action be sufficiently evident to the user?
- Will the user notice that the correct action is available?
- Will the user associate and interpret the response from the action correctly?

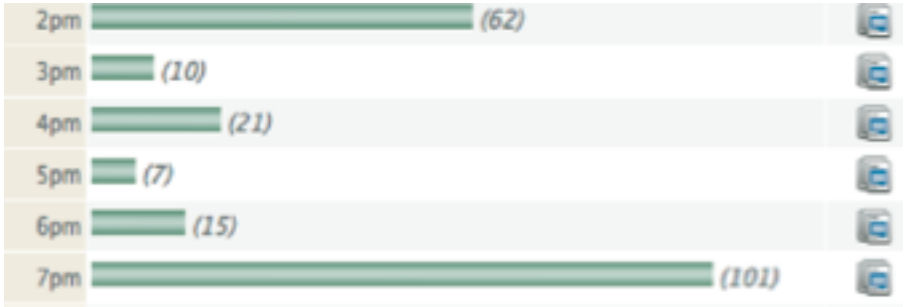
As the experts work through the scenario they note problems.

Pluralistic walkthrough

- Variation on the cognitive walkthrough theme.
- Performed by a carefully managed team.
- The panel of experts begins by working separately.
- Then there is managed discussion that leads to agreed decisions.
- The approach lends itself well to participatory design.

Analytics

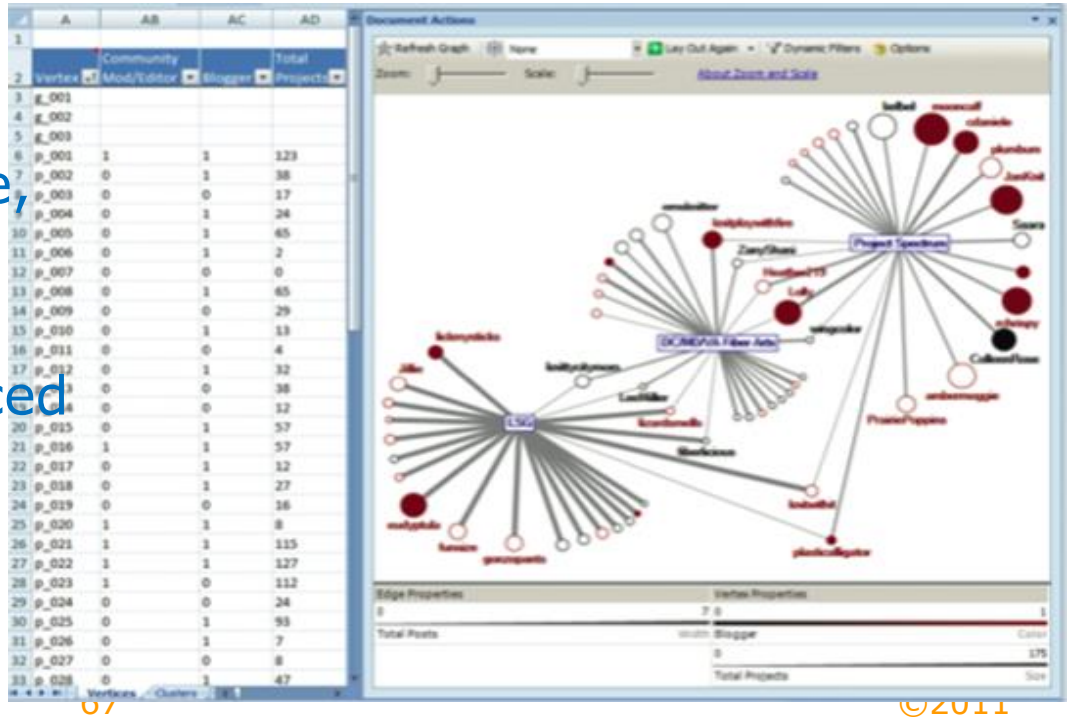
- A method for evaluating user traffic through a system or part of a system
- Many examples including Google Analytics, Visistat (shown below)
- Times of day & visitor IP addresses



Social network analysis tools
(Perer & Shneiderman, 2008)

Perer A and Shneiderman B (2008), "Systematic yet flexible discovery: guiding domain experts through exploratory data analysis", In Proceedings of the 13th International Conference on Intelligent User Interfaces. New York, NY, USA , pp. 109-118. ACM.

- Focus on relationships rather than individual elements
- How individuals interact and the influence they have on each other.
- Find pattern in the structure, flow of resources through a network
- How individuals are influenced by their surroundings



Predictive models

- Provide a way of evaluating products or designs without directly involving users.
- Less expensive than user testing.
- Usefulness limited to systems with predictable tasks - e.g., telephone answering systems, mobiles, cell phones, etc.
- Based on expert error-free behavior.

GOMS

- Goals – what the user wants to achieve eg. find a website.
- Operators - the cognitive processes & physical actions needed to attain goals, eg. decide which search engine to use.
- Methods - the procedures to accomplish the goals, eg. drag mouse over field, type in keywords, press the go button.
- Selection rules - decide which method to select when there is more than one.

Keystroke level model

- GOMS has also been developed to provide a quantitative model - the keystroke level model.
- The keystroke model allows predictions to be made about how long it takes an expert user to perform a task.

Response times for keystroke level operators (Card et al., 1983)

Operator	Description	Time (sec)
K	Pressing a single key or button	
	Average skilled typist (55 wpm)	0.22
	Average non-skilled typist (40 wpm)	0.28
	Pressing shift or control key	0.08
	Typist unfamiliar with the keyboard	1.20
P	Pointing with a mouse or other device on a display to select an object. This value is derived from Fitts' Law which is discussed below.	0.40
P1	Clicking the mouse or similar device	0.20
H	Bring 'home' hands on the keyboard or other device	0.40
M	Mentally prepare/respond	1.35
R(t)	The response time is counted only if it causes the user to wait.	t

Summing together

$$T_{\text{execute}} = T_{\text{K}} + T_{\text{P}} + T_{\text{H}} + T_{\text{D}} + T_{\text{M}} + T_{\text{R}}$$

Using KLM to calculate time to change gaze (Holleis et al., 2007)

Fitts' Law (Fitts, 1954)

- Fitts' Law predicts that the time to point at an object using a device is a function of the distance from the target object & the object's size.
- The further away & the smaller the object, the longer the time to locate it & point to it.
- Fitts' Law is useful for evaluating systems for which the time to locate an object is important, e.g., a cell phone, a handheld devices.

Key points

- Inspections can be used to evaluate requirements, mockups, functional prototypes, or systems.
- User testing & heuristic evaluation may reveal different usability problems.
- Walkthroughs are focused so are suitable for evaluating small parts of a product.
- Analytics involves collecting data about users activity on a website or product
- The GOMS and KLM models and Fitts' Law can be used to predict expert, error-free performance for certain kinds of tasks.