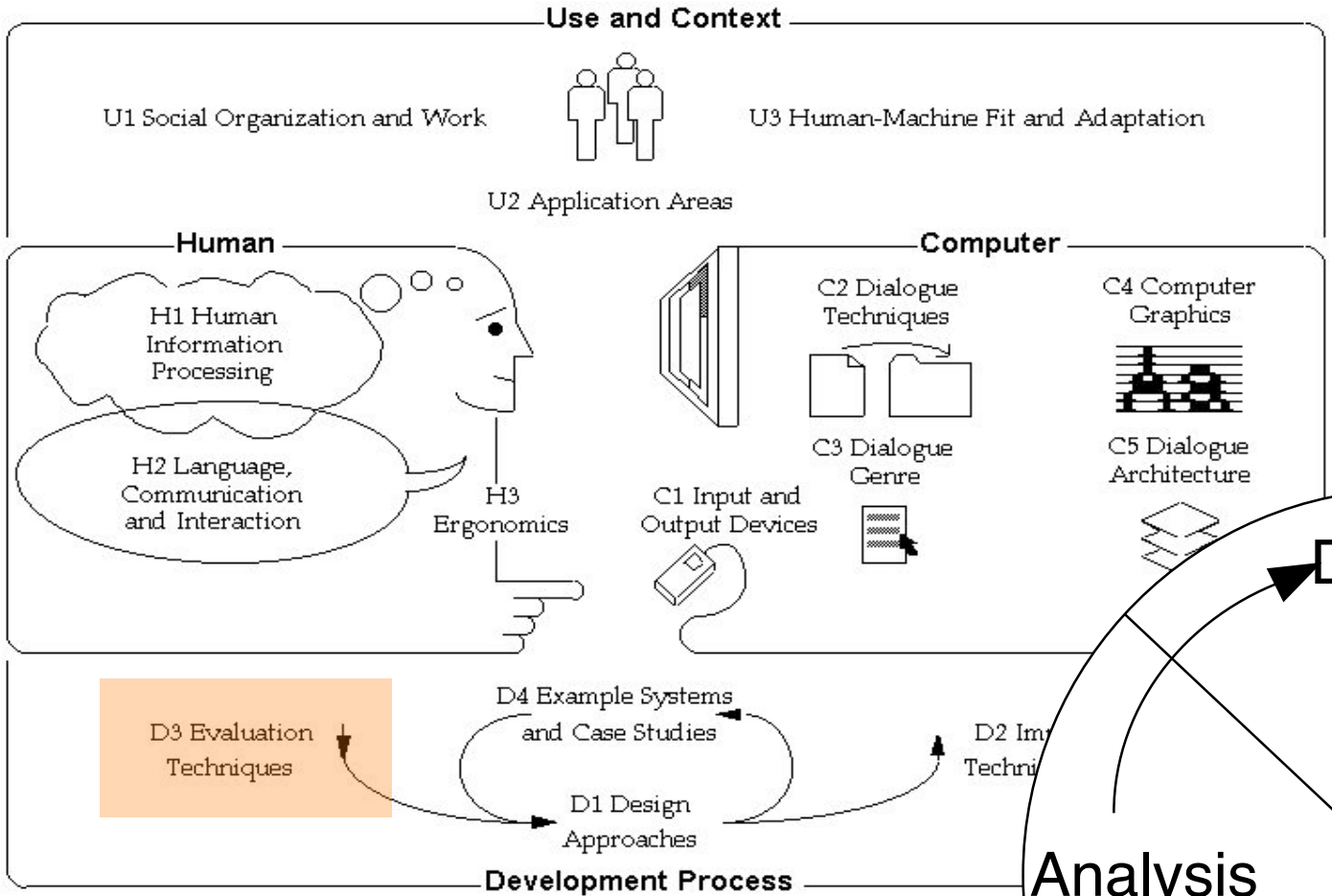
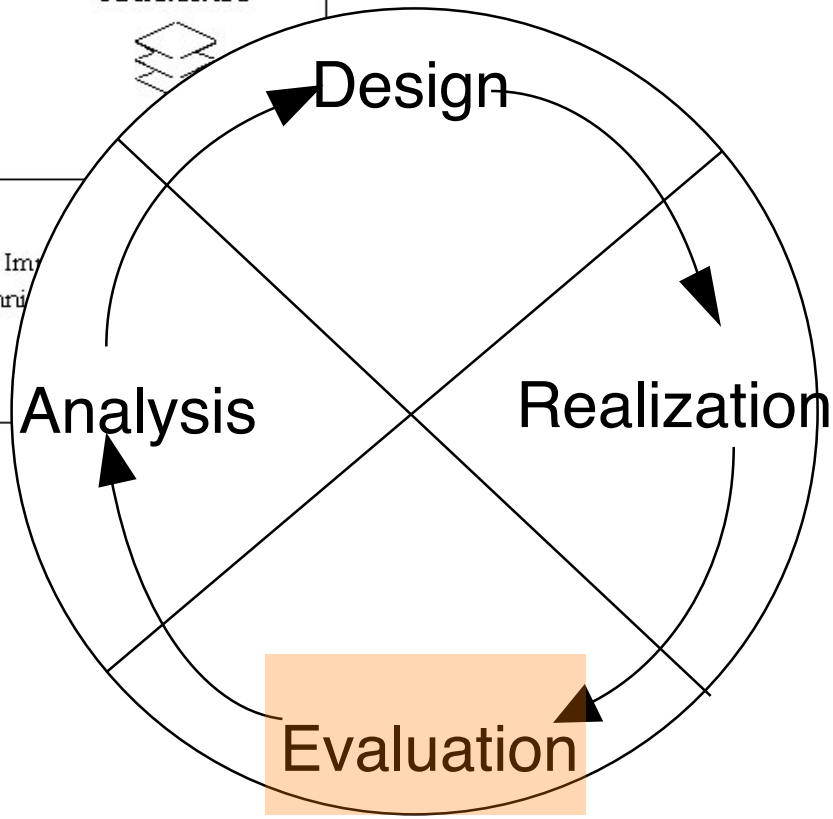
8 Evaluation

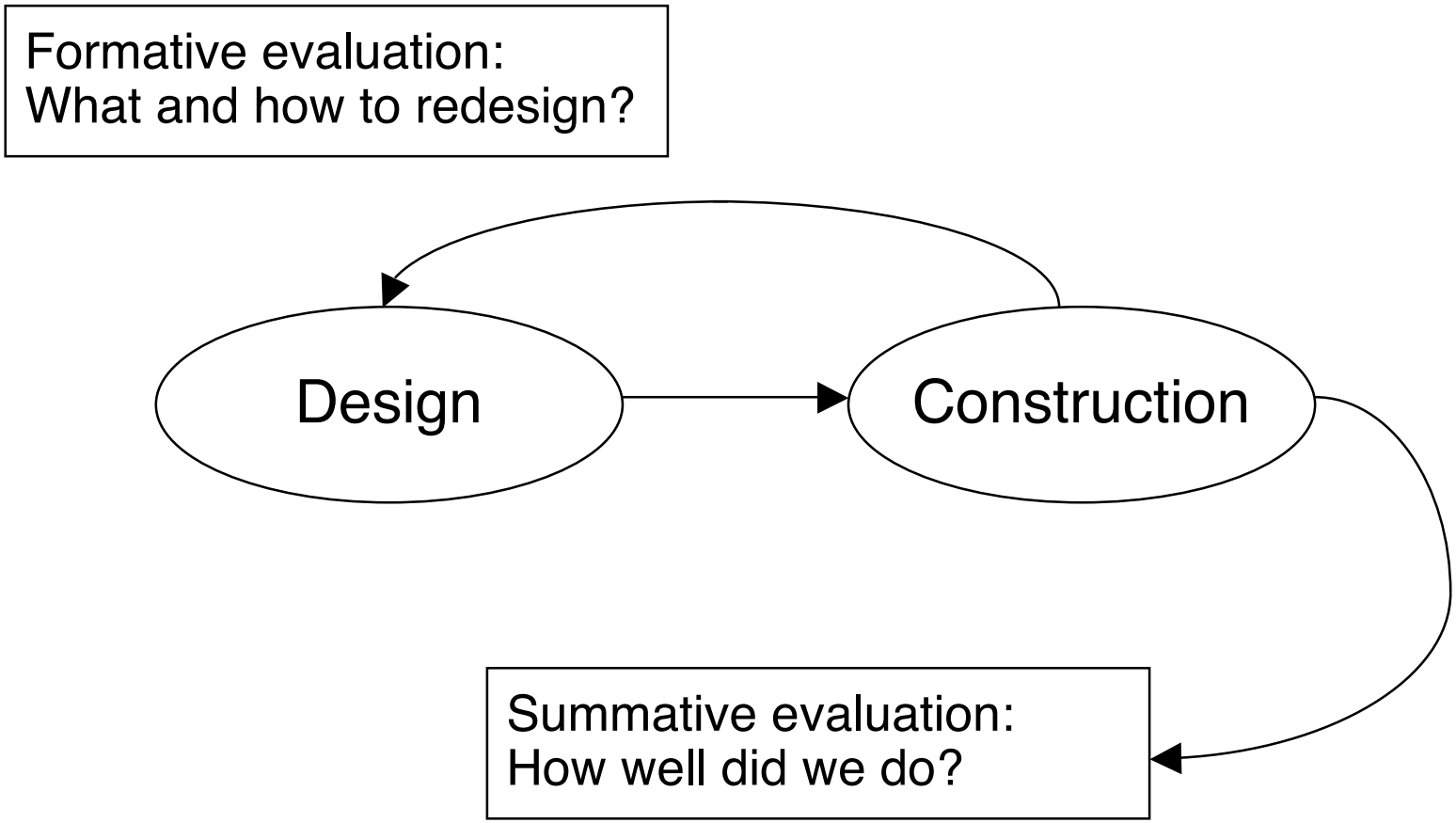

Use and Context
U1 Social Organization and Work
U3 Human-Machine Fit and Adaptation
U2 Application Areas
Human
Computer
H1 Human Information Processing
H2 Language, Communication and Interaction
H3 Ergonomics
C1 Input and Output Devices
C2 Dialogue Techniques
C3 Dialogue Genre
C4 Computer Graphics
C5 Dialogue Architecture
D3 Evaluation Techniques
D4 Example Systems and Case Studies
D1 Design Approaches
Development Process


Design
Analysis
Realization
Evaluation

What to evaluate?

- The usability of a system!
- … it depends on the stage of a project
  - Ideas and concepts
  - Designs
  - Prototypes
  - Implementations
  - Products in use
- … it also depends on the goals
- Approaches
  - Formative vs. summative evaluation
  - Analytical vs. empirical evaluation
  - Qualitative vs. quantitative results

Formative vs. Summative Evaluation

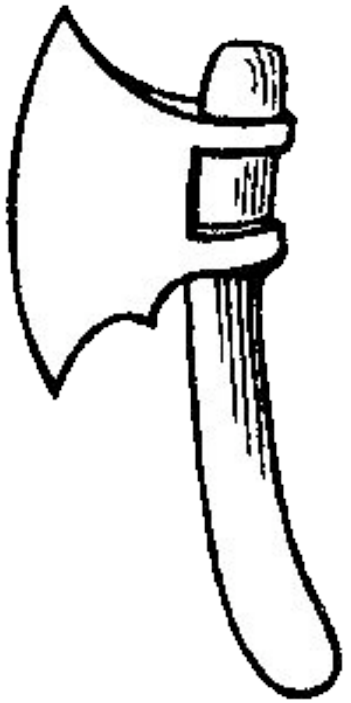


- M. Scriven: The methodology of evaluation, 1967

Analytic vs. Empirical Evaluation

- Scriven, 1967:

  “If you want to evaluate a tool, say an axe, you might study the design of the bit, the weight distribution, the steel alloy used, the grade of hickory in the handle, etc., or you may just study the kind and speed of the cuts it makes in the hands of a good axeman.”



Empirical and Analytic Methods are Complementary

- Empirical evaluation produces facts which need to be interpreted
  - If the axe does not cut well, what do we have to change?
  - Analytic evaluation identifies the crucial characteristics
- Analytic evaluation produces facts which need to be interpreted
  - Why does the axe have a special-shaped handle?
  - Empirical evaluation helps to understand the context for object properties

Orthogonality of Approaches (Examples)

	Formative	Summative
Analytic	• Tools for usability assessment • Cognitive walkthroughs	• Heuristic evaluation • Standards compliance evaluation
Empirical	• User group identification • Prototype user study	• Usability lab test • Field studies

Evaluation without Criteria is Useless

- Possible criteria (examples):
  - Informal assessment of one idea against another
  - Detailed statistical analysis of average performance using realistic user group (or actual field usage)
  - Fulfilment of informal usability heuristics
  - Fulfilment of formalized usability-related design metrics
  - …
- In any case:
  - We have to know in advance what we are looking for before we can evaluate!

Usability Methods are often not used!

- Why
  - Developers are not aware of it
  - The expertise to do evaluation is not available
  - People do not know about the range of methods available
  - Certain methods are too expensive for a project
    - » or people think they are too expensive
  - Developers see no need because the product “works”
  - Teams think their informal methods are good enough

8 Evaluation

Types of Analytic Evaluation

- Inspection-based evaluation
  - Expert review
  - Heuristic evaluation
  - Cognitive walkthrough
- Model-based evaluation
  - Evaluation of design models
  - Derivation of model views
- Different results
  - Qualitative assessment
  - Quantitative assessment

Inspections & Expert Review

- Throughout the development process
- Performed by developers and experts
- External or internal experts
- Tool for finding problems
- May take between an hour and a week
- Structured approach is advisable
  - Reviewers should be able to communicate all their issues (without hurting the team)
  - Reviews must not be offensive for developers / designers
  - The main purpose is finding problems
  - Solutions may be suggested but decisions are up to the team

Inspection Methods

- Guideline review
  - Check that the UI is according to a given set of guidelines
- Consistency inspection
  - Check that the UI is consistent (in itself, within a set of related applications, with the OS)
  - Bird's eye view can help
    - » e.g. printout of a web site and put it up on the wall)
  - Consistency can be enforced by design (e.g. CSS for Web sites)
- Procedure for inspections:
  - Find reviewers, define schedule
  - Prepare material for reviewers, including criteria
  - On-site or off-site review
  - Review report, definition of consequences

Informal Evaluation

- Expert reviews and inspections are often done informally
  - UIs and interaction is discussed with colleagues
  - People are asked to comment, report problems, and suggest additions
  - Experts (often within the team) assess the UI for conformance with guidelines and consistency
- Results of informal reviews and inspections are often directly used to change the product
  - … still state of the art in many companies!
  - The personal view of the CEO, or his partner …
- Really helpful evaluation
  - Is explicit
  - Has clearly documented findings
  - Can increase the quality significantly
- Expert reviews and inspections are a starting point for change

Discount Usability Engineering (Nielsen 1994)

- Low cost approach
- Small number of subjects
  - First users often give most valuable feedback
  - Developers, team members may act as subjects
- Approximate
  - Get indications and hints
  - Find major problems
  - Discover many issues (minor problems)
- Qualitative approach
  - Observe user interactions
  - User explanations and opinions
  - Anecdotes, transcripts, problem areas, …
- Quantitative approach
  - Count, log, measure something of interest in user actions
  - Speed, error rate, counts of activities

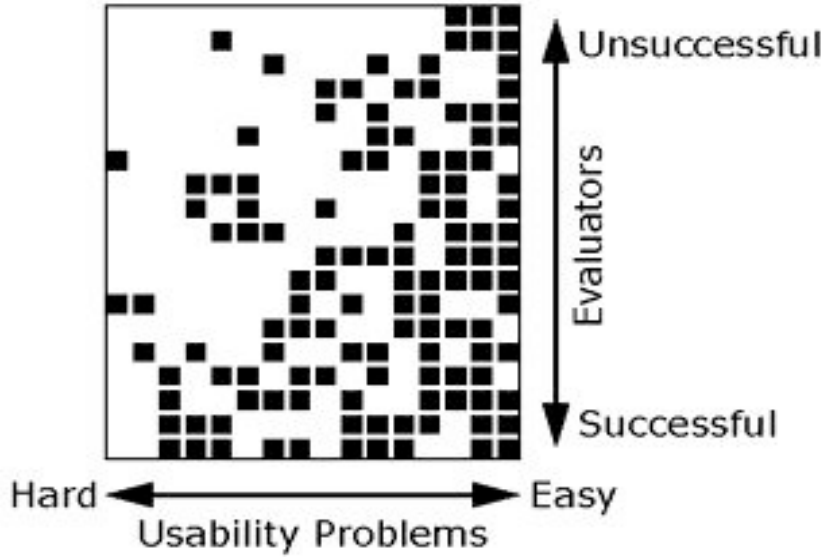
http://www.useit.com/papers/guerrilla_hci.html

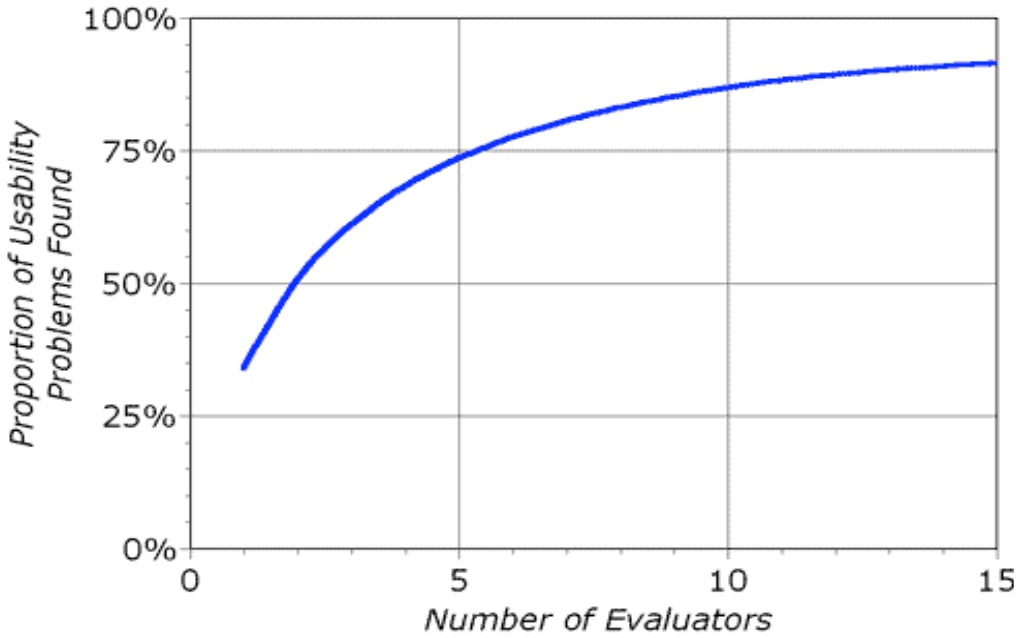
Heuristic Evaluation

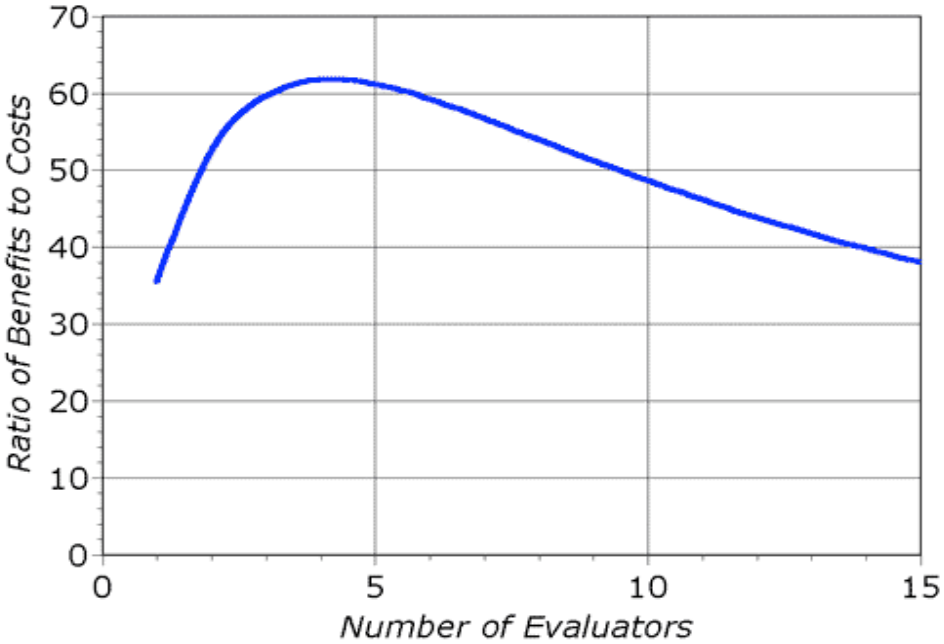
- Heuristic evaluation is a “discount” usability inspection method
  - Quick, cheap and easy evaluation of UI design
  - http://www.useit.com/papers/heuristic/
- Basic Idea:
  - Small set of evaluators examine the interface and judge its compliance with recognized usability principles (the "heuristics").
    - » Either just by inspection or by scenario-based walkthrough
    - » Critical issues list, weighted by severity grade
    - » Opinions of evaluators are consolidated into one report
- Extremely popular:
  - Google search for “heuristic evaluation”
    - » 1998: 600 pages
    - » 2002: 9.500 pages
    - » 2007: 213.000 pages
- Implicit assumptions:
  - There exists a fixed list of desirable properties of user interfaces (the “heuristics”)
  - These heuristics can be checked by experts with a clear and defined result

Number of Evaluators

- How many evaluators?
- Example: total cost estimate with 11 evaluators at about 105 hours, see
  http://www.useit.com/papers/guerrilla_hci.html







Ten Usability Heuristics (Nielsen)

- Visibility of system status
- Match between system and the real world
- User control and freedom
- Consistency and standards
- Error prevention
- Recognition rather than recall
- Flexibility and efficiency of use
- Aesthetic and minimalist design
- Help users recognize, diagnose, and recover from errors
- Help and documentation

Detailed Checklist Example (1)

Usability Techniques
Heuristic Evaluation - A System Checklist

By Deniese Pierotti, Xerox Corporation

http://www.stcsig.org/usability/topics/articles/he-checklist.html

Heuristic Evaluation - A System Checklist

1. Visibility of System Status

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

#	Review Checklist	Yes No N/A	Comments
1.1	Does every display begin with a title or header that describes screen contents?	O O O	
1.2	Is there a consistent icon design scheme and stylistic treatment across the system?	O O O	
1.3	Is a single, selected icon clearly visible when surrounded by unselected icons?	O O O	
1.4	Do menu instructions, prompts, and error messages appear in the same place(s) on each menu?	O O O	
1.5	In multipage data entry screens, is each page labeled to show its relation to others?	O O O	
1.6	If overtype and insert mode are both available, is there a visible indication of which one the user is in?	O O O	
1.7	If pop-up windows are used to display error messages, do they allow the user to see the field in error?	O O O	
1.8	Is there some form of system feedback for every operator action?	O O O	
1.9	After the user completes an action (or group of actions), does the feedback indicate that the next group of actions can be started?	O O O	
1.10	Is there visual feedback in menus or dialog boxes about which choices are selectable?	O O O	
1.11	Is there visual feedback in menus or dialog boxes about which choice the cursor is on now?	O O O	

Detailed Checklist Example (2)

2. Match Between System and the Real World

The system should speak the user's language, with words, phrases and concepts familiar to the user, rather than system-orient conventions, making information appear in a natural and logical order.

#	Review Checklist	Yes No N/A	Comments
2.1	Are icons concrete and familiar?	O O O	
2.2	Are menu choices ordered in the most logical way, given the user, the item names, and the task variables?	O O O	
2.3	If there is a natural sequence to menu choices, has it been used?	O O O	
2.4	Do related and interdependent fields appear on the same screen?	O O O	
2.5	If shape is used as a visual cue, does it match cultural conventions?	O O O	
2.6	Do the selected colors correspond to common expectations about color codes?	O O O	
2.7	When prompts imply a necessary action, are the words in the message consistent with that action?	O O O	
2.8	Do keystroke references in prompts match actual key names?	O O O	
2.9	On data entry screens, are tasks described in terminology familiar to users?	O O O	
2.10	Are field-level prompts provided for data entry screens?		
2.11	For question and answer interfaces, are questions stated in clear, simple language?	O O O	
2.12	Do menu choices fit logically into categories that have readily understood meanings?	O O O	
2.13	Are menu titles parallel grammatically?	O O O	
2.14	Does the command language employ user jargon and avoid computer jargon?	O O O	

Tognazzini's First Principles

- Anticipation
- Autonomy of the user
- Color Blindness
- Consistency
- Inconsistency
- Defaults
- Efficiency of the user
- Explorable Interfaces
- Fitt's law
- Human-interface objects (standard, consistent, stable)
- Latency reduction
- Learnability
- Metaphors, use of
- Protect user's work
- Readability
- Track State
- Visible Interfaces

http://www.stanford.edu/group/web-creators/heuristics

Problems with Inspection Methods

- Validity of the findings:
  “Usability checklists and inspections can produce rapid feedback, but may call attention to problems that are infrequent or atypical in real-worlds use.” (Rosson/Carroll)
- Usage context for inspection
  - Selection of scenarios, or decision not to use scenarios, may influence results heavily
- Systematic contribution to the discipline of usability engineering?
  - Heuristic evaluation relies very much on creativity and experience of the evaluators
  - How to save and reuse the knowledge available in the heads of expert evaluators?

Cognitive Walkthrough

- One or more evaluators going through a set of tasks
  - Evaluating understandability and ease of learning
- Procedure:
  - Defining the input:
    - » Who will be the users of the system?
    - » What task(s) will be analyzed?
    - » What is the correct action sequence for each task?
    - » How is the interface defined?
  - During the walkthrough:
    - » Will the users try to achieve the right effect?
    - » Will the user notice that the correct action is available?
    - » Will the user associate the correct action with the effect to be achieved?
    - » If the correct action is performed, will the user see that progress is being made toward solution of the task?

From www.usabilityhome.com

Model-Based Evaluation

- Models of interactive applications
  - Wide range of syntactic and semantic frameworks
  - Sometimes models are available early as part of specifications
- Examples:
  - GOMS (Card et al.), see Chapter 2
  - User Action Notation (UAN) (Siochi/Hartson/Hix)
  - Petri Nets and other Formal Methods
  - UML-based approaches, see Chapter 6
  - ConcurTaskTrees (CTT)
- Basic idea:
  - Usability analysis carried out on an abstraction of the actual system
  - Potential for automation of guideline-based evaluation
- Key problem:
  - Where to get the model from? Very low level of models!
  - Specification? Reverse engineering of code into model?

UAN Example

Task: **select file**		
USER ACTIONS	**INTERFACE FEEDBACK**	**INTERFACE STATE**
~[file icon'] Mv	file icon'! ¥file icon * file icon' : file icon-!	selected = file

Not

Not equal

Highlights

For all

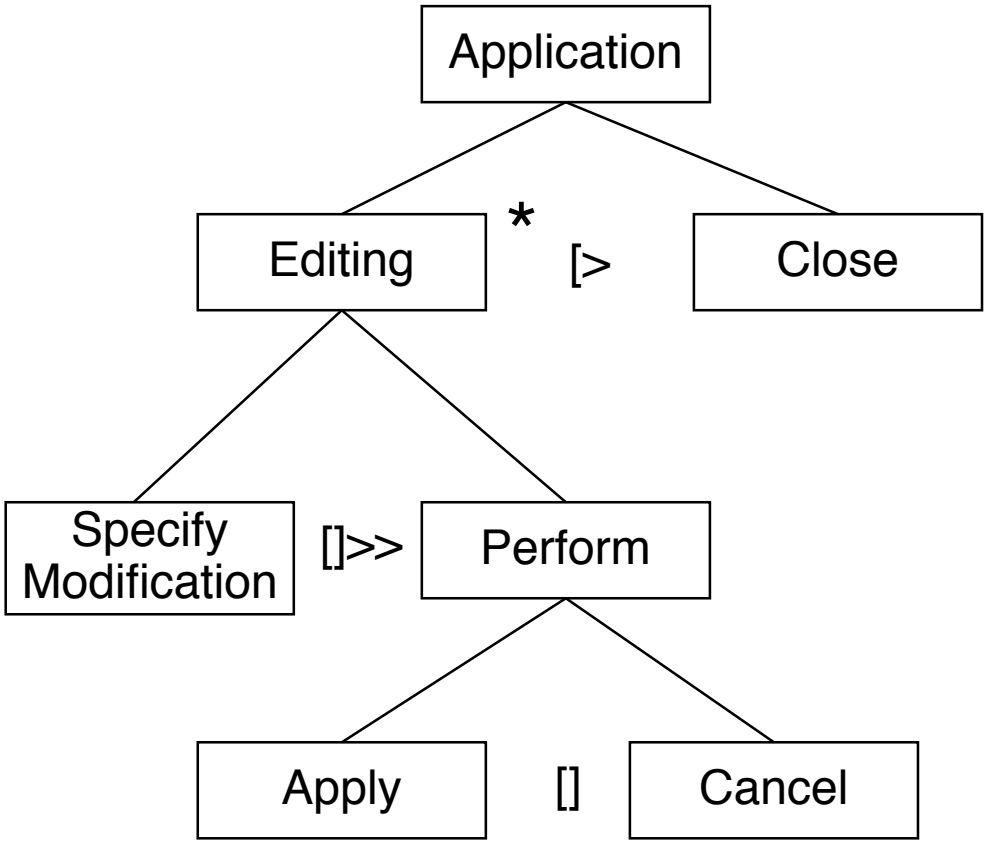
Down

Mouse

Move cursor

http://www.it.bton.ac.uk/staff/rng/teaching/notes/UAN.html

CTT Example



Interaction task

[> Deactivation (of first task by second)

[]>> Enabling with information passing

[] Choice

Notation based on LOTOS (Language of Temporal Ordering Specification), in turn based on process algebra languages

Fabio Paternó

8 Evaluation

Types of Empirical Evaluation

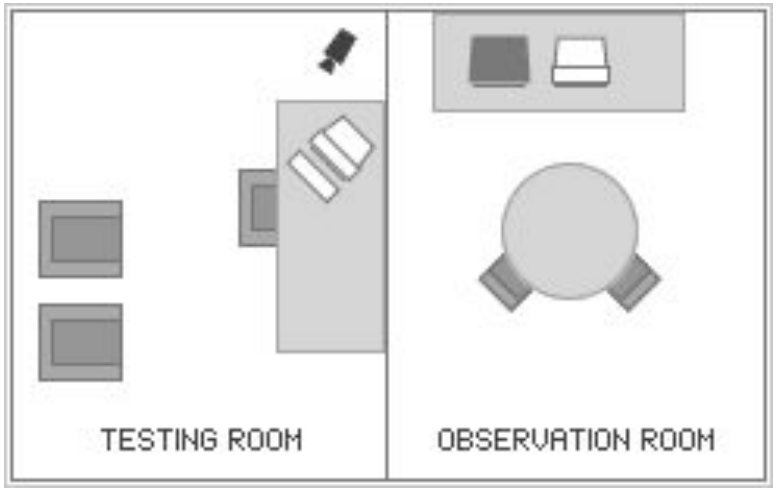
- Field Studies
  - Short-term studies
  - Longitudinal studies
- Laboratory Usability Testing
  - General usability testing
  - Benchmark testing
- Controlled Experiments
  - Quantitative user studies
  - (see tutorials)
- Physiological measurement
  - E.g. eye-tracking
  - (see Chapter 2)

Field Studies

- Normal activities are studied in normal environment
- Advantages:
  - Can reveal results on user acceptance
  - Allows longitudinal studies, including learning and adaptation
- Problems:
  - In general very expensive
  - Highly reliable product (prototype, mockup) needed
  - How to get observations?
    - » Collecting usage data
    - » Collecting incident stories
    - » On-line feedback
    - » Retrospective interviews, questionnaires

Usability Laboratory

- Specifically constructed testing room
  - Instrumented with data collection devices (e.g. microphones, cameras)
- Separate observation room
  - Usually connected to testing room by one-way mirror and audio system
  - Data recording and analysis
- Test users perform prepared scenarios
  - “Think aloud” technique
- Problem:
  - Very artificial setting
  - No communication



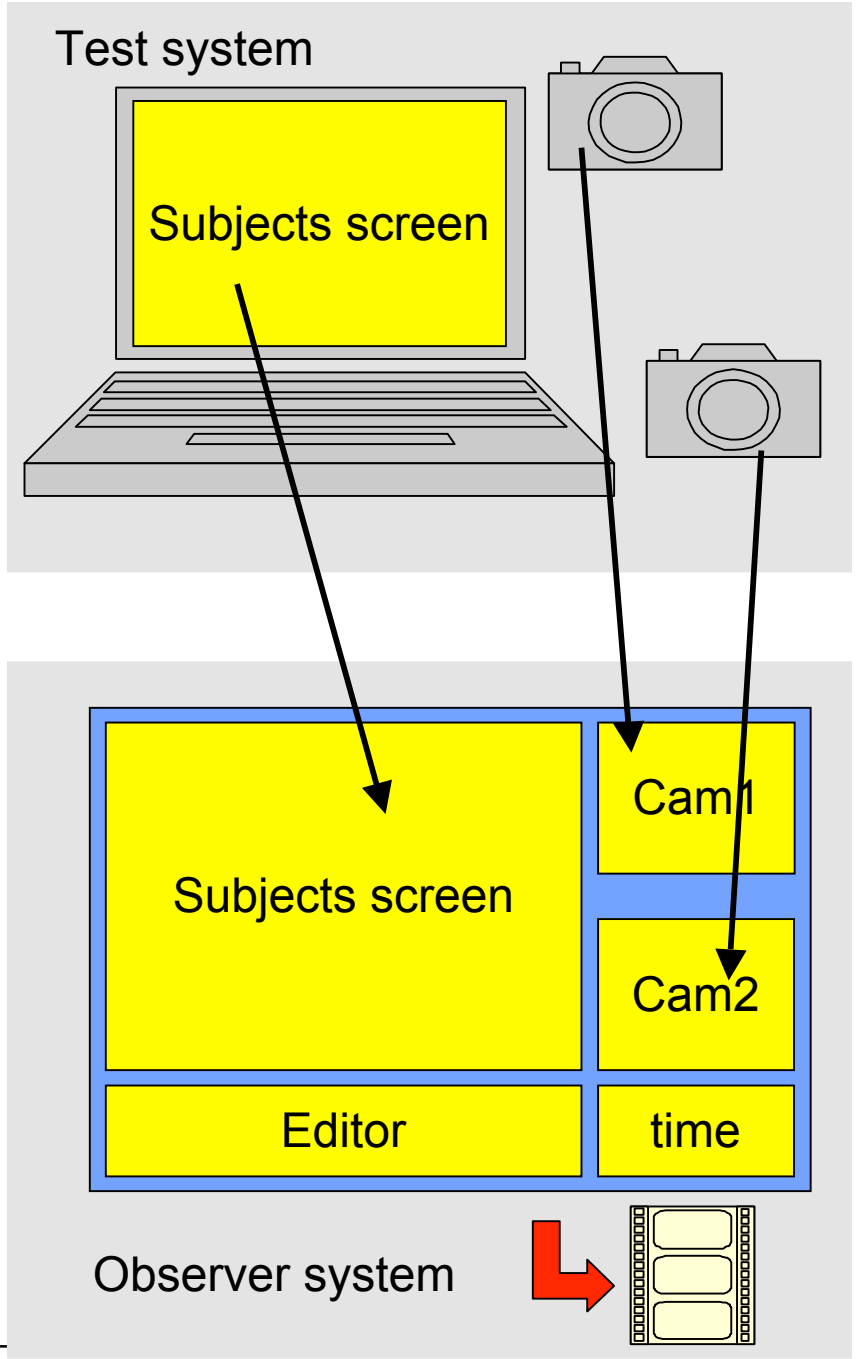
Source: www.xperienceconsulting.com

Video protocol

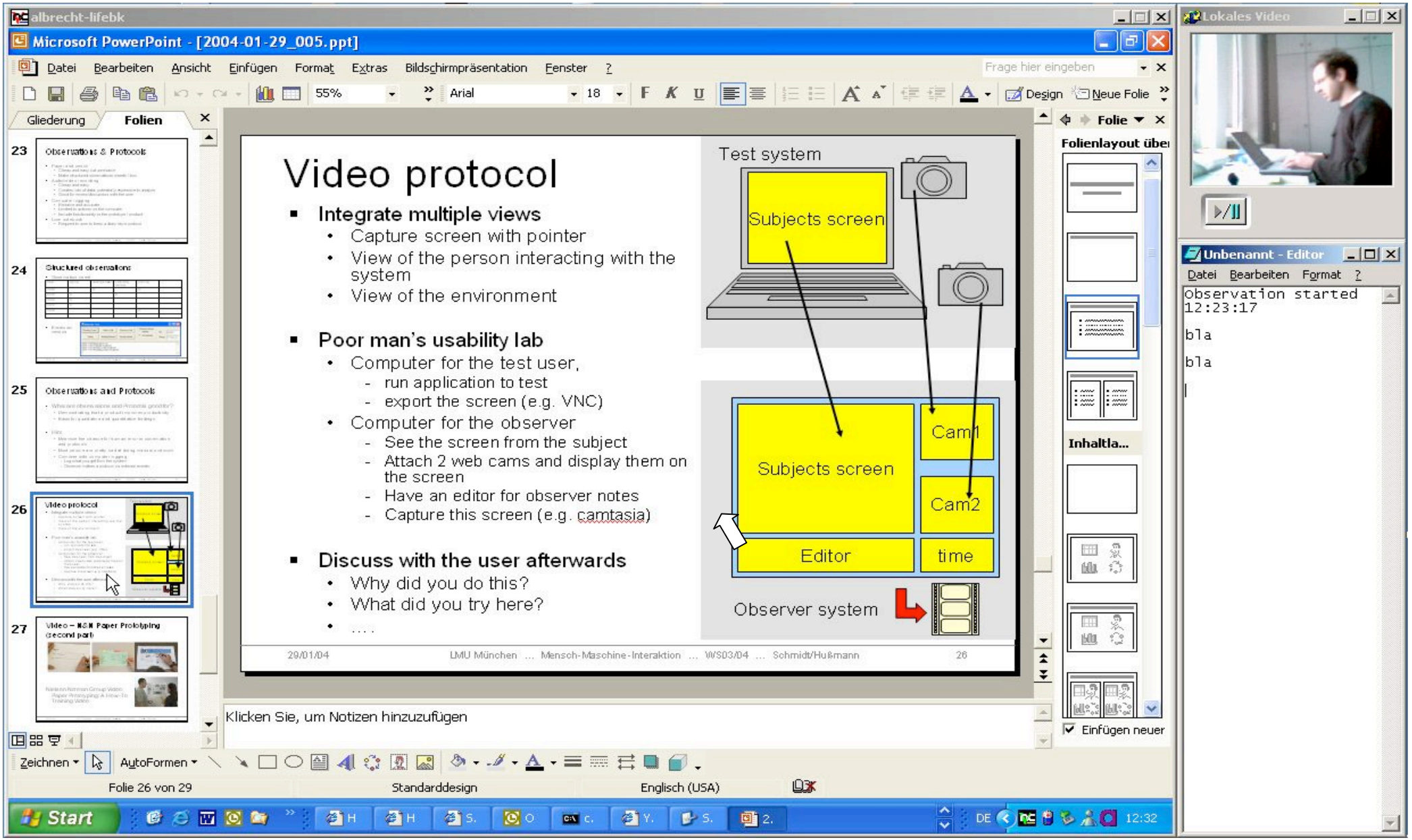
- Integrate multiple views
  - Capture screen with pointer
  - View of the person interacting with the system
  - View of the environment

- Poor man's usability lab
  - Computer for the test user,
    - » run application to test
    - » export the screen (e.g. VNC)
  - Computer for the observer
    - » See the screen from the subject
    - » Attach 2 web cams and display them on the screen
    - » Have an editor for observer notes
    - » Capture this screen (e.g. Camtasia)

- Discuss with the user afterwards
  - Why did you do this?
  - What did you try here?
  - ….



Screen video



Controlled Experiments

- Answering specific additional, often quantitative, questions
  - Performance
  - Satisfaction
- Providing basic knowledge generic to many applications
  - Comparing input/output devices
  - Comparing general design strategies
- Basic idea:
  - Selected participants carry out well-defined tasks
  - Specific values (variables) are measured and compared
- Principal experiment designs:
  - Within-subjects design:
    - » Same participant exposed to all test conditions
  - Between-subjects design:
    - » Independent groups of participants for each test condition

Variables in Experiment Design

- Variables are manipulated and measured
  - ***Independent*** variables are manipulated
  - ***Dependent*** variables are measured
- The conditions of the experiment are set by independent variables
  - E.g. number of items in a list, text size, font, color
  - The number of different values used is called *level*
  - The number of experimental conditions is the product of the levels
  - E.g. font can be times or arial (2 levels), background can be blue, green, or white (3 levels). This results in 6 experimental conditions (times on blue, times, on green, …, arial on white)
- The dependent variables are the values that can be measured
  - Objective values: e.g. time to complete a task, number of errors, etc.
  - Subjective values: ease of use, preferred option
  - They should only be dependent on changes of the independent variables

Hypotheses

- Prediction of the result of an experiment
- Stating how a change in the independent variables will effect the measured dependent variables
- With the experiment it can be shown that the hypothesis is correct
- Usual approach
  - Stating a null-hypothesis (predicts that there is no effect)
  - Carrying out the experiment and using statistical measures to disprove the null-hypothesis
  - When a statistical test shows a significant difference it is probable that the effect is not random
- Carefully apply statistical significance tests
  - (see tutorials)

Example:
Study on Text Input

- Is text input by keyboard really better than using T9 on a phone?

- Compare text input speed and errors made
  - Qwertz-keyboard on a notebook computer
  - T9 on a mobile phone

- Concentrate on test input only, ignore:
  - Time to setup / boot / initialize the device
  - Time to get into the application

Example:
Study on Text Input (2)

- Participants
  - How many?
  - Skills
    - » Computer user?
    - » Phone/T9 users?

- Independent variables
  - Input method
  - Text to input
- Dependent variables
  - Time to input a text
  - Number of errors made

Example:
Study on Text Input (3)

- Independent variables
  - Input method,
    - » 2 levels: Keyboard and T9
  - Text to input
    - » 1 level: text with about 10 words

- Experimental conditions
  - 2 conditions – T9 and Key
  - User 1,3,5,7,9 perform T9 then Key
  - User 2,4,6,8,10 perform Key then T9
  - Different texts in first and second run?
  - Particular phone model?
  - Completion time is measured (e.g. stop watch or application)
  - Number of errors/corrections is observed

Example:
Study on Text Input (4)

- Hypotheses
  - H-1: Input by keyboard is quicker than T9
  - H-2: fewer errors are made using keyboard input compared to T9

- Null-Hypotheses
  - Assumes no effect
  - H0-1: there is no difference in the input speed between keyboard and T9
  - H0-2: there is no difference in the number of errors made using a keyboard input compared to T9

8 Evaluation

Classification of Analytic Evaluation Techniques

	Cognitive walkthrough	Heuristic evaluation	Review based	Model based
Stage	Throughout	Throughout	Design	Design
Style	Lab	Lab	Lab	Lab
Objective?	No	No	As source	No
Measure	Qualitative	Qualitative	As source	Qualitative
Information	Low level	High level	As source	Low level
Immediacy	N/a	N/a	As source	N/a
Intrusive?	No	No	No	No
Time	Medium	Low	Low-medium	Medium
Equipment	Low	Low	Low	Low
Expertise	High	Medium	Low	High

Classification of Experimental Evaluation Techniques

	Experiment	Think aloud	Post-task walkthrough	Physiological measurement
Stage	Throughout	Implementn.	Implementn.	Implementn.
Style	Lab	Lab/field	Lab/field	Lab
Objective?	Yes	No	No	Yes
Measure	Quantitative	Qualitative	Qualitative	Quantitative
Information	Low/high level	High/low level	High/low level	Low level
Immediacy	Yes	Yes	No	Yes
Intrusive?	Yes	Yes	No	Yes
Time	High	High	Medium	Medium/high
Equipment	Medium	High	Low	High
Expertise	Medium	High	Medium	High

References Chapter 7

- Alan Dix, Janet Finlay, Gregory Abowd and Russell Beale: Human Computer Interaction (third edition), Prentice Hall 2003
- Mary Beth Rosson, John M. Carroll: Usability Engineering. Morgan-Kaufman 2002. Chapter 7
- Discount Usability Engineering http://www.useit.com/papers/guerrilla_hci.html
- Heuristic Evaluation http://www.useit.com/papers/heuristic/