Vorlesung Mensch-Maschine-Interaktion

Albrecht Schmidt
Embedded Interaction Research Group
LFE Medieninformatik
Ludwig-Maximilians-Universität München
http://www.hcilab.org/albrecht/

Chapter 7
Evaluation
(selected topics)

- **7.1 User studies**
- 7.2 Heuristic 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 evaluation – throughout the design, helps to shape a product
  - Summative evaluation – quality assurance of the finished product.

Why Studies and Experiments?

- To measure more precisely the usability or other features
- Applicable mainly to
  - Functional prototypes
  - Testing an implementation
  - Quality monitoring of software products
- To compare solutions, e.g.
  - Users are quicker using version A than using version B
  - Users make 10% less errors when using version X than when using version Y
- To provide quantitative figures, e.g.
  - 90% of the users can complete the transaction using version Y in less than 3 minutes
  - On average users will be able to by a ticket using version A in less than 30 seconds

Designing the experiment

- Basic Scientific Method
  1. Form Hypothesis
  2. Collect data
  3. Analyze
  4. Accept/reject hypothesis

- Issues for user studies
  - System to test
  - Participants
  - Hypothesis
  - Variables
  - Experimental Methods
  - Statistical approach

Does computer science fit this traditional science approach?

Is it really possible to prove usability?

Procedure for user studies

- Set goals
- Design the experiment
- Schedule users
- For each user (typical example):
  - Inform the user and sign the consent form
  - Do a survey on demographics and questions of interest to the experiment
  - Give the participant instructions on the task – do not reveal the hypotheses
  - (optional) Make a training run - depends on the study
  - Perform the actual run and measure variables
  - (optional) do a survey on subjective measure
  - Be available for questions of participants or for their (informal) feedback
- Analyze the results

Participants (Subjects)

- How many participants do we need?
  - Depending on the project and the goals
  - Depending on the set-up
    - measuring the login-in time of remote users vs.
    - Doing a full video observation for a 1 hour task
  - Be pragmatic
  - Minimal size of about 10 participants
- Participants should be representative for the user group
  - Age, background, skills, experience, …
  - In most cases the other people on the team are NOT representative!
- How to recruit participants
  - Customer data base
  - Market research services
  - Volunteers (online, newspapers, etc) – this is risky because the people who will respond are often not representative

Selection of Participants

- Services offered that allow to get participants fitting a specific description
- Methods widely used in market research

- Example: Online Panel
  - For online questionnaires
  - Pool of users
  - Customer can specify the users that should take part

- How do companies get their subjects?
  - Incentive (money, prices, …)
  - Big set of questions when registering potential users, show examples from ComCult Online Panel

Variables

- 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 hypotheses is correct
- Usual approach
  - Stating a null-hypotheses (this predicts that there is not effect of the change in the independent variable on the measured variable)
  - Carrying out the experiment and using statistical measures to disprove the null-hypotheses
  - When a statistical test shows a significant difference it is probable that the effect is not random

Designing the experiment

- The experiment should be set up to be reproducible!
- Main factors
  - Participants
  - Independent variables
  - Hypotheses stated
- Approach
  - state the hypotheses – what do you want to proof
  - find the variables? Which are varied? which are measured?
  - Find participants – representative for the experiment
  - Fix the method to use (between-groups / within groups)

Experimental Method

- Within groups
  - Each user performs under all the different conditions
  - Important to randomize the order of the conditions for each participant
  - Problems
    - Learning may influence results
  - Advantages
    - The effect of differences between individuals are lessened
    - Fewer participants required

- Between groups (randomize)
  - One condition is selected for each participant
  - Each user performs only under one condition (avoids learning)
  - Careful selection of groups is essential
  - Drawback
    - Differences between individuals in different groups can play an important role (leads to large groups)
    - More user required
    - Usually harder to show significance

Statistical Tests

- See statistics text book (e.g. form psychology or medical tests)
- Software packages offer functions
- Test selected depends on
  - Distribution of the measured variables
  - The type of variables (continuous or discrete)
  - Experimental Method
- Example: Student's t-test
  - On the difference of means
  - Assumes a normal distribution
  - Functions available in spreadsheet software and statistics packages
- Example ANOVA
  - Analysis of Variance

- "significant difference"
  - Simplified: the probability that effect observed is random is less the 0.05

T-Test example in Excel

- TTEST(…)
- Parameters
  - Data row 1
  - Data row 2
  - Ends (1 or 2)
  - Type (paired, same variance, different variance)

User	Time M1	Time M2		
100	37	31		
101	44	38		
102	42	43		
103	56	37		
104	99	50		
105	33	30		
106	45	50		
107	49	36		
108	70	71		
109	63	56		
110	54	51		
111	61	46		
average	54,4167	44,9167		
t test (paired)			0,042	TTEST(B7:B18;C7:C18;2;1)
t test (un-paired)			0,137	TTEST(B7:B18;C7:C18;2;2)

Further Issues

- Consent form – get written consent from participants
  - Templates available
  - May be checked with the legal department / review board
- Let participants know what they are doing
  - What is the participant expected to do
  - Procedure
  - How long will it take, breaks
  - What is the study for in general – but do NOT tell about the specific purpose or your hypotheses
- Make sure they know
  - Quality of a UI / software is tested
  - They are NOT tested
- Ethical Issues

Participants Consent (Example)

Participants Consent Form

Study _____ **Institution** _____

Name: _____ Date of Birth: _____
Email: _____
Phone:_____

I have been informed on the procedure and purpose of the study and my questions have been answer to my satisfaction.

I have volunteered to take part in this study and agree that during the study information is recorded (audio and video as well as my interaction with the system). This information may only be used for research and teaching purpose. I understand that my participation in this study is confidential. All personal information and individual results will not be released to third parties without my written consent.

I understand that I can withdraw from participation in the study at any time.

Date: _____ Signature:_____

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 than Key
  - User 2,4,6,8,10 perform Key than T9
  - Different texts in first and second run?
  - Particular phone model?
  - Completion time is measure (e.g. stop watch or application)
  - Number of error/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
- Experimental Method
  - Within groups
  - Randomized order of conditions

Example: Study on Text Input (5)

- Collect Data

User	Order	Time Cond1	Time Cond2	# Err Cond1	# Err Cond2
01	c1>c2	…	…	…	…
02	c2>c1	…	…	…	…
03	c1>c2	…	…	…	…

- Perform a statistical analysis
- … exercise on Friday.

Example:
Study on Text Input (6)

- Fairness
  - Same conditions and procedure (e.g. light condition, interruptions, noise)
  - Specify procedure for exceptions (e.g. someone does not complete the test)
  - No bias
- Participants Consent
- Further Issues?
  - Ethical issues
  - Privacy

Chapter 7
Evaluation
(selected topics)

- 7.1 User studies
- **7.2 Heuristic 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 evaluation – throughout the design, helps to shape a product
  - Summative evaluation – quality assurance of the finished product.

Why evaluate?
Goals of user interface evaluation

- Ensure functionality (effectiveness)
  - Assess (proof) that a certain task can be performed
- Ensure performance (efficiency)
  - Assess (proof) that a certain task can be performed given specific limitations (e.g. time, resources)
- Customer / User acceptance
  - What is the effect on the user?
  - Are the expectations met?
- Identify problems
  - For specific tasks
  - For specific users
- Improve development life-cycle
- Secure the investment (don't develop a product that can only be used by fraction of the target group – or not at all!)

There is not a single way …

- Different approaches
  - Inspections
  - Model extraction
  - Controlled studies
  - Experiments
  - Observations
  - Field trails
  - Usage context

- Different results
  - Qualitative assessment
  - Quantitative assessment

Usability Methods are often not used!

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

- starting points
  - Discount Usability Engineering
    http://www.useit.com/papers/guerrilla_hci.html
  - Heuristic Evaluation
    http://www.useit.com/papers/heuristic/

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 and Expert Review 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)
  - Birds'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 on the web)
- Walkthrough
  - Performing specific tasks (as the user would do them)
- Heuristic evaluation
  - Check that the UI violates a set (usually less than 10 point) rules

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!
- Informal evaluation is important but in most cases not enough

- Making evaluation more explicit and documenting the findings can increase the quality significantly
- Expert reviews and inspections are a starting point for change

Discount Usability Engineering

- Low cost approach
- Small number of 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

Heuristic Evaluation

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

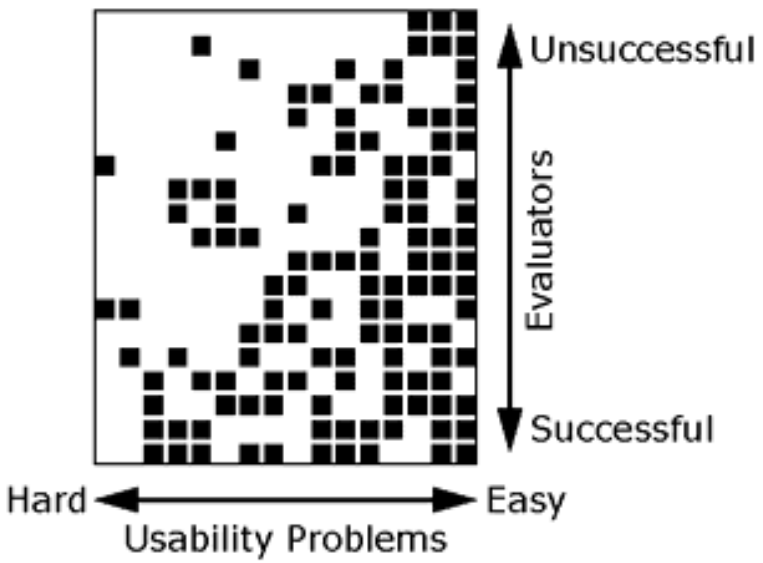
- Heuristic evaluation is a usability inspection method
- systematic inspection of a user interface design for usability
- goal of heuristic evaluation
  - to find the usability problems in the design
- As part of an iterative design process.

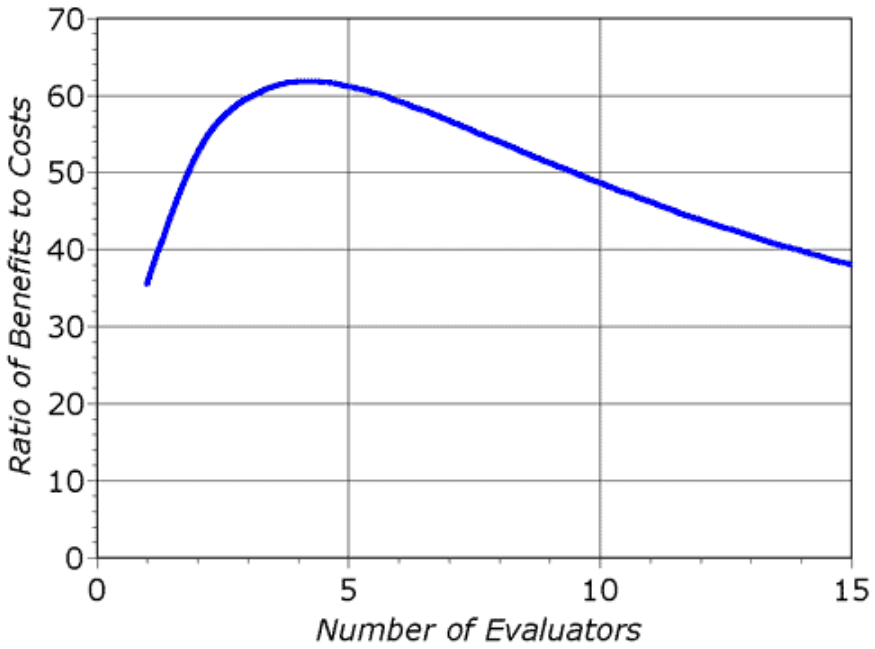
- Basic Idea:
  Small set of evaluators examine the interface and judge its compliance with recognized usability principles (the "heuristics").

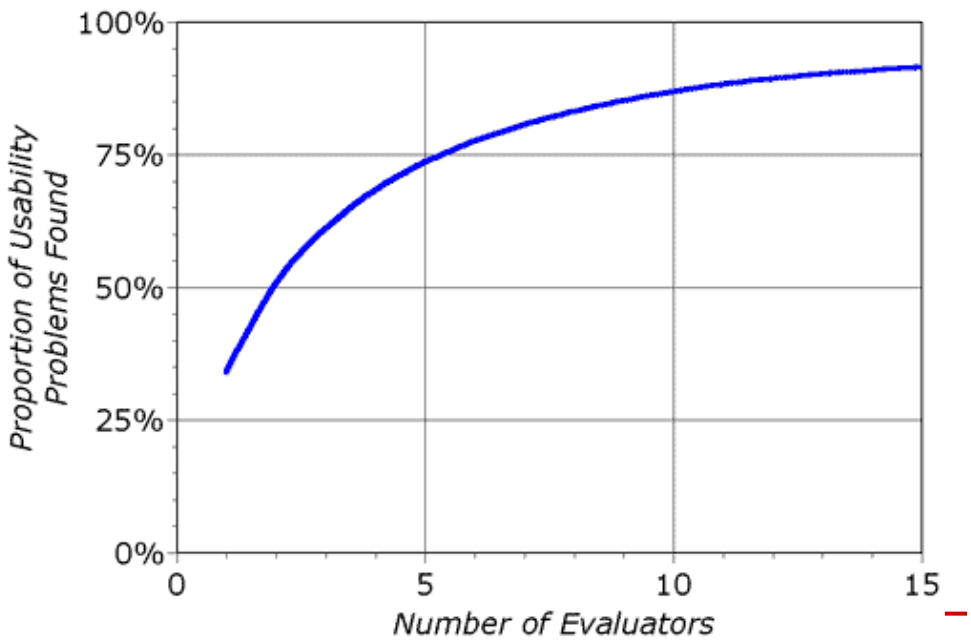
Heuristic Evaluation

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

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







Heuristic Evaluation - Heuristics

- Heuristics suggested by 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

- Depending of the product and goals a different set may be appropriate

Heuristic Evaluation - Steps

- Preparation
  - Assessing appropriate ways to use heuristic evaluation
  - Define Heuristics
  - Having outside evaluation expert learn about the domain and scenario
  - Finding and scheduling evaluators
  - Preparing the briefing
  - Preparing scenario for the evaluators
  - Briefing (system expert, evaluation expert, evaluators)
  - Preparing the prototype (software/hardware platform) for the evaluation
- Evaluation
  - Evaluation of the system by all evaluators
  - Observing the evaluation sessions
- Analysis
  - Debriefing (evaluators, developers, evaluation expert)
  - compiling list of usability problems (using notes from evaluation sessions)
  - Writing problem descriptions for use in severity-rating questionnaire
  - Severity rating

Heuristic Evaluation – Severity Rating

- Severity ratings are used to prioritize problems
- Decision whether to release a system or to do further iterations
- The severity of a usability problem is a combination of three factors:
  - The frequency with which the problem occurs: Is it common or rare?
  - The impact of the problem if it occurs: Will it be easy or difficult for the users to overcome?
  - The persistence of the problem: Is it a one-time problem that users can overcome once they know about it or will users repeatedly be bothered by the problem

- 0 to 4 rating scale to rate the severity of usability problems:
  - 0 = I don't agree that this is a usability problem at all
  - 1 = Cosmetic problem only: need not be fixed unless extra time is available on project
  - 2 = Minor usability problem: fixing this should be given low priority
  - 3 = Major usability problem: important to fix, so should be given high priority
  - 4 = Usability catastrophe: imperative to fix this before product can be released

Observations & Protocols

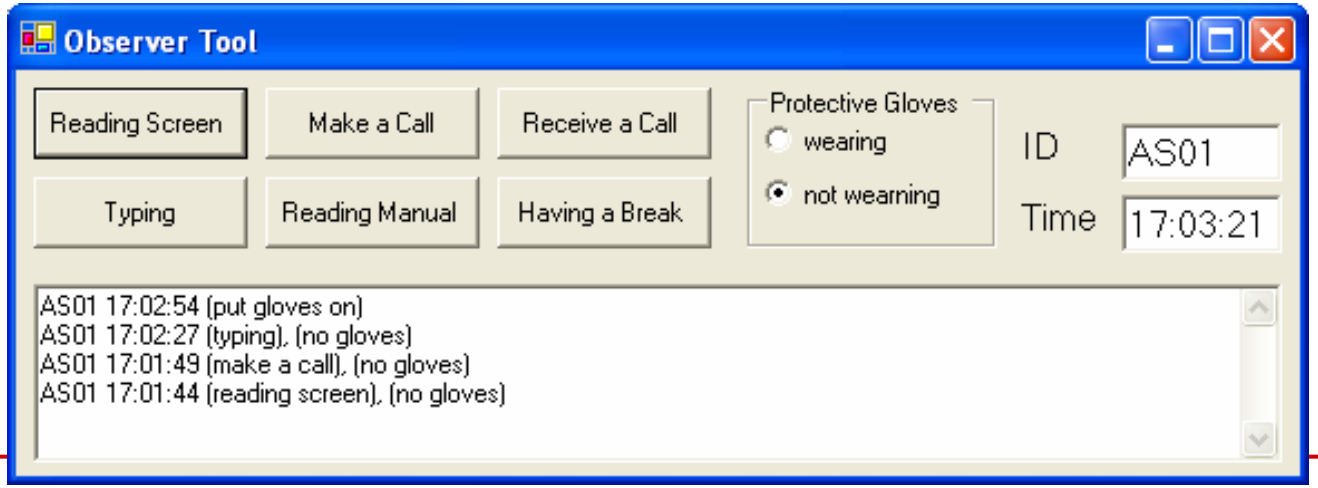
- Paper and pencil
  - Cheap and easy but unreliable
  - Make structured observations sheets / tool
- Audio/video recording
  - Cheap and easy
  - Creates lots of data, potentially expensive to analyze
  - Good for review/discussion with the user
- Computer logging
  - Reliable and accurate
  - Limited to actions on the computer
  - Include functionality in the prototype / product
- User notebook
  - Request to user to keep a diary style protocol

Structured observations

- Observation sheet

time	typing	reading screen	consulting manual	phoning	…
14:00		X		X	
14:01	X		X		
14:02	X				
14:03	X				
14:04				X	
…					

- Electronic version



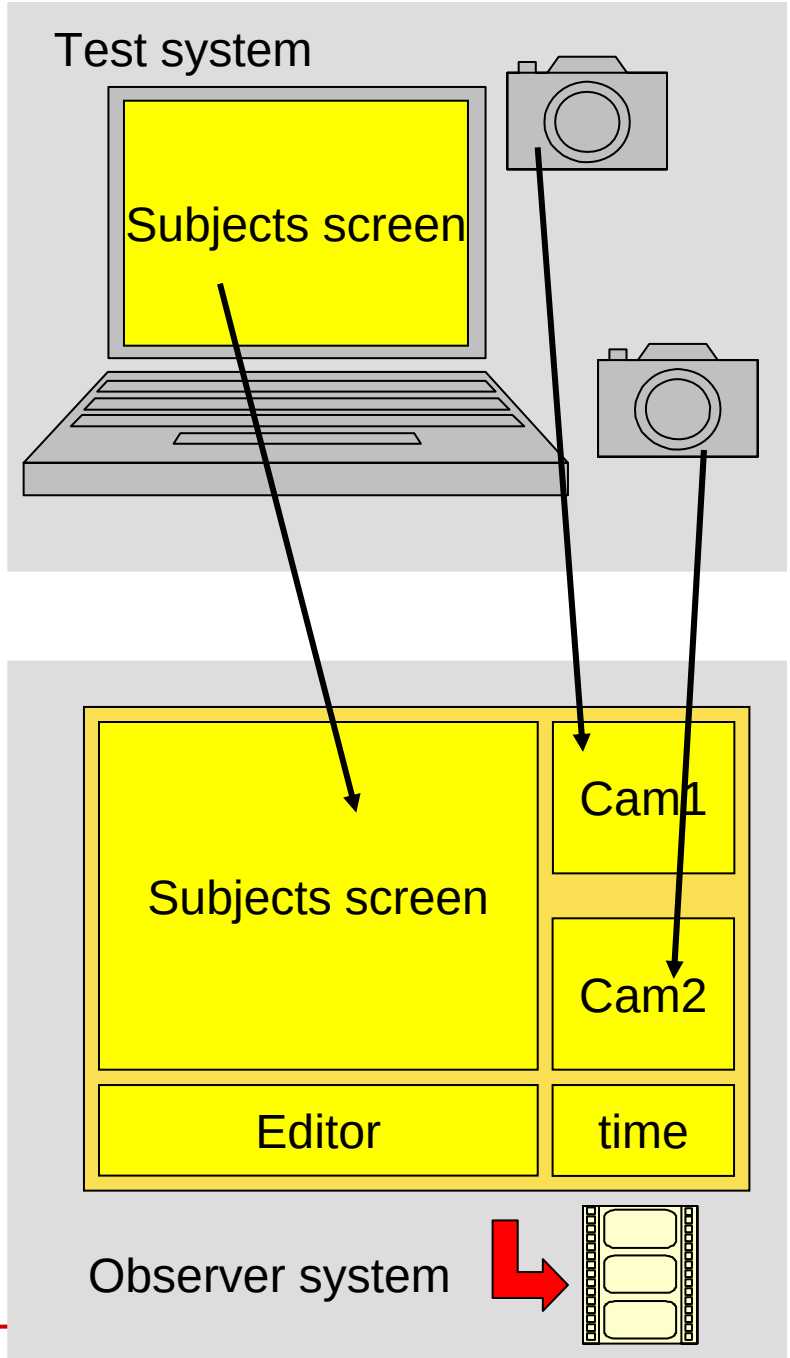
Observations and Protocols

- What are observations and Protocols good for?
  - Demonstrating that a product improves productivity
  - Basis for qualitative and quantitative findings

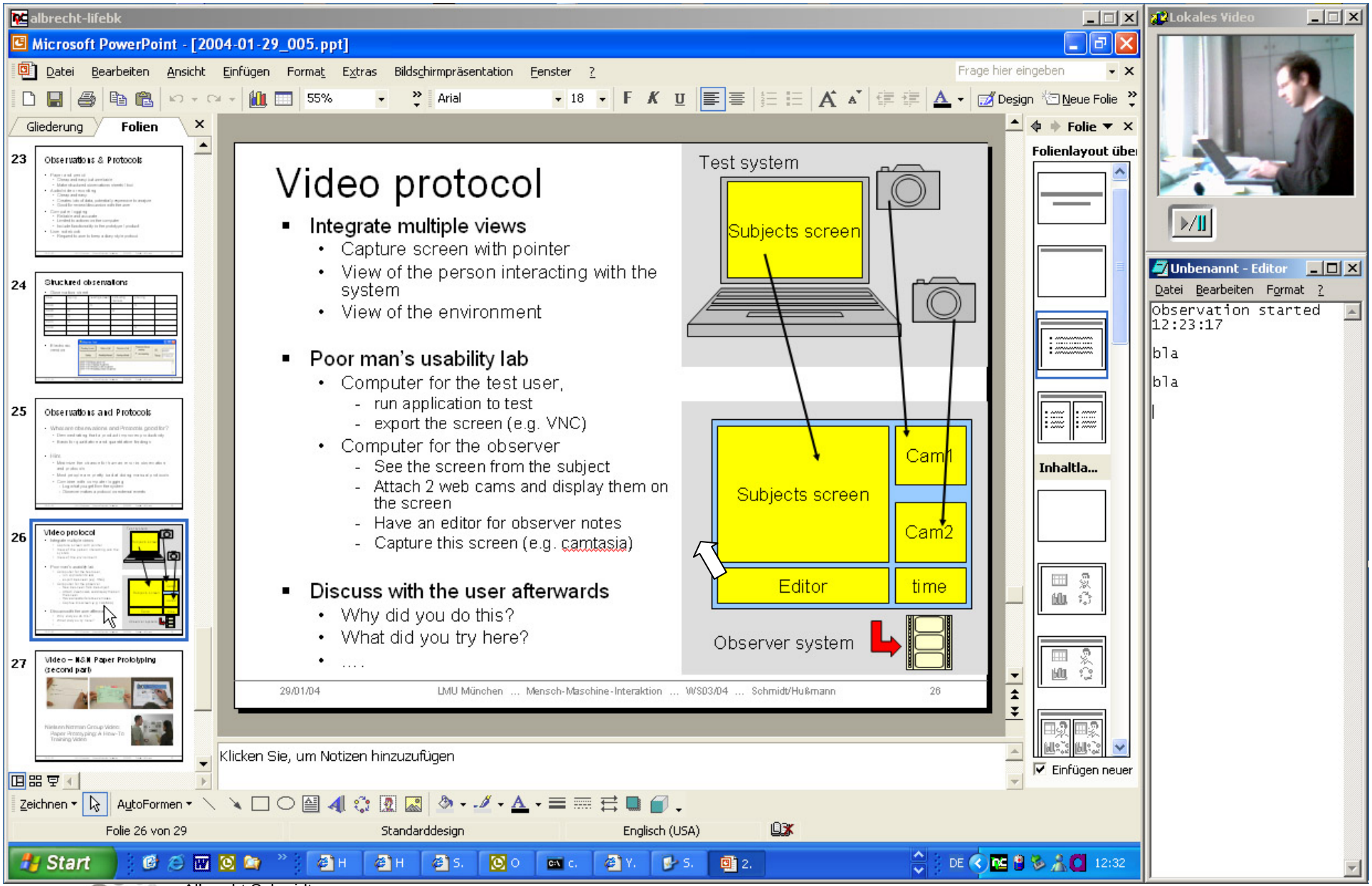
- Hint
  - Minimize the chance for human error in observation and protocols
  - Most people are pretty bad at doing manual protocols
  - Combine with computer logging
    - Log what you get from the system
    - Observer makes a protocol on external events

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



References Chapter 7

- Alan Dix, Janet Finlay, Gregory Abowd and Russell Beale. (1998) Human Computer, Interaction (second edition), Prentice Hall, ISBN 0132398648 (new Edition announced for October 2003)
- Ben Shneiderman. (1998) Designing the User Interface, 3rd Ed., Addison Wesley; ISBN: 0201694972

- Discount Usability Engineering http://www.useit.com/papers/guerrilla_hci.html
- Heuristic Evaluation http://www.useit.com/papers/heuristic/