Evaluation and Testing

Andreas Butz, LMU Media Informatics
butz@lmu.de
slides partially taken from MMI class

What and when to evaluate

- Original title: Evaluation and user testing
  - ...we‘ll talk about testing products, not users, right?
  - general mindset: accept the **user as the reference**!

- „Evaluation“ somehow sounds as if it only happens at the end
  - evaluation methods can be used throughout the entire development process!

Tests	Development	Tests	Process	Tests		Test

The user as the ultima ratio...

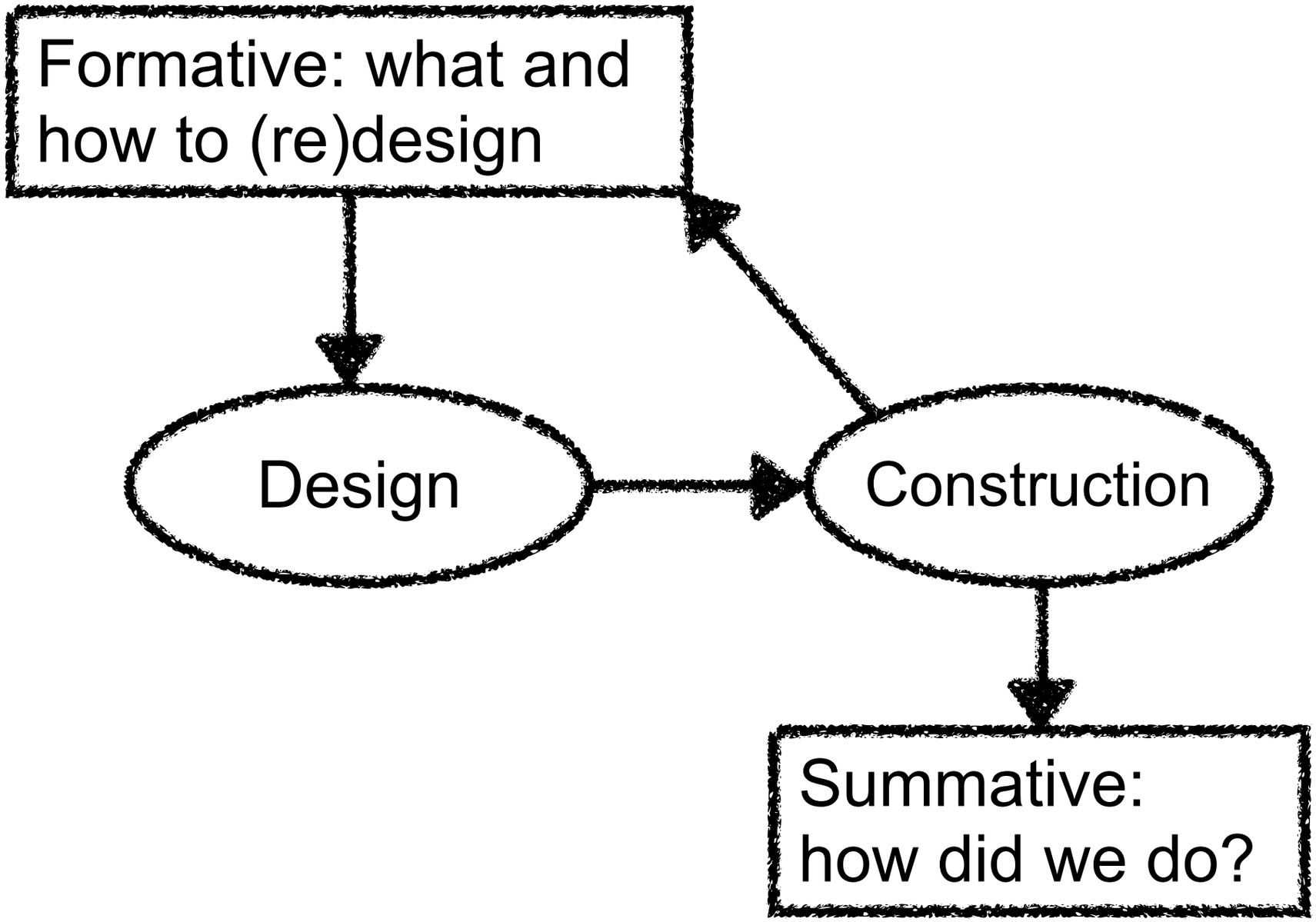
Donald Norman

What can be evaluated?

- The usability of a system!
- … it depends on the stage of a project
  - Ideas and concepts
  - Designs
  - (paper and functional) Prototypes
  - Implementations
  - Products in use

- … it also depends on the goals
- Approaches:
  - Formative vs. summative evaluation
  - Analytical vs. empirical evaluation
  - Qualitative vs. quantitative results

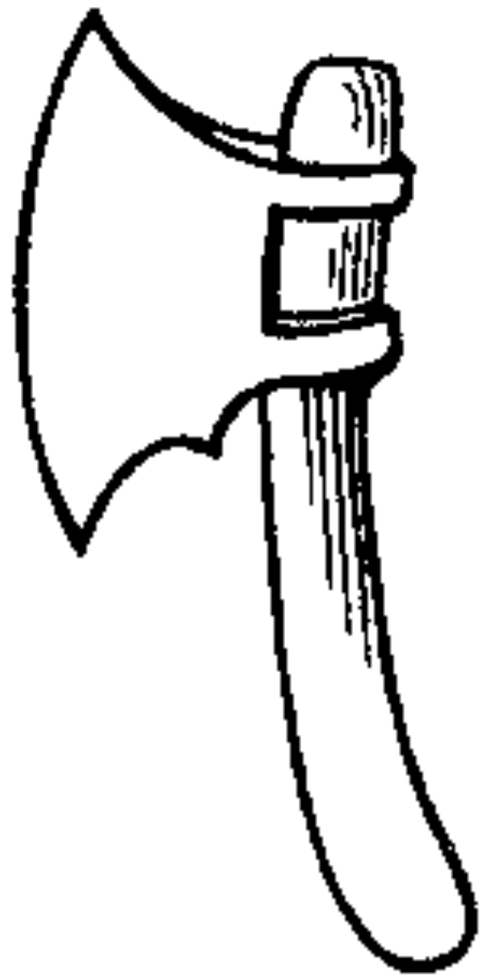
Formative vs. Summative Evaluation



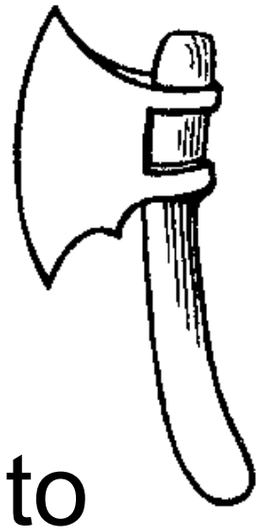
- M. Scriven: The methodology of evaluation, 1967

Analytical 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 (not complimentary ;-)



- 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

- Analytical 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

Agenda for this class

Intro & motivation		
	Formative	Summative
Analytical	Cognitive walkthrough	Heuristic evaluation
Empirical	Prototype user study	Controlled experiment Usability lab test Field studies
Discussion and take-home thoughts		

Agenda for this class

Intro & motivation		
	Formative	Summative
Analytical	Cognitive walkthrough	Heuristic evaluation
Empirical	Prototype user study	Controlled experiment Usability lab test Field studies
Discussion and take-home thoughts		

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

Agenda for this class

Intro & motivation		
	Formative	Summative
Analytical	Cognitive walkthrough	Heuristic evaluation
Empirical	Prototype user study	Controlled experiment Usability lab test Field studies
Discussion and take-home thoughts		

Heuristic Evaluation


Jakob Nielsen

- 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
- Implicit assumptions:
  - There is a fixed list of desirable properties of UIs (the “heuristics”)
  - They can be checked by experts with a clear and defined result

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

Usability Techniques
Heuristic Evaluation - A System Checklist

By Deniese Pierotti, Xerox Corporation

Heuristic Evaluation - A System Checklist http://www.stcsig.org/usability/topics/articles/he-checklist.html

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	

Problems with Inspection Methods

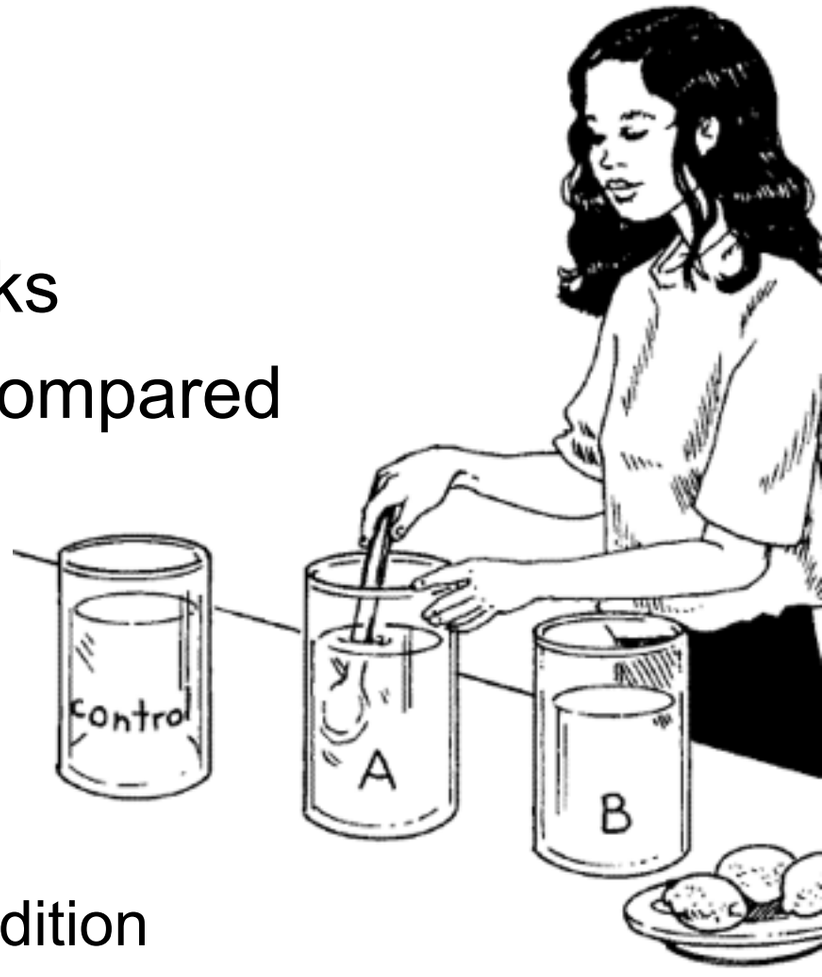
- Validity of the findings:
  “Usability checklists and inspections can produce rapid feedback, but may call attention to problems that are infrequent or atypical in real world use.” (Rosson/Carroll)
- Usage context for inspection
  - Selection of scenarios, or decision not to use scenarios, may influence 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?

Agenda for this class

Intro & motivation		
	Formative	Summative
Analytical	Cognitive walkthrough	Heuristic evaluation
Empirical	Prototype user study	Controlled experiment Usability lab test Field studies
Discussion and take-home thoughts		

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



http://01.edu-cdn.com/files/static/wiley/9780471419204/CONTROLLED_EXPERIMENTS_01.GIF

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 affect the measured dependent variables
- With the experiment it can be tested whether 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 statistics lecture!)

Example: Study on Text Input

- Is text input on a 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

- Participants
  - How many?
  - Skills
    - Computer user?
    - Phone/T9 users?
- Independent variables
  - Input method,
    - 2 levels: Keyboard and T9
  - Text to input
    - 1 level: text with about 10 words
- Dependent variables
  - Time to input the text
  - Number of errors made

Example: Study on Text Input

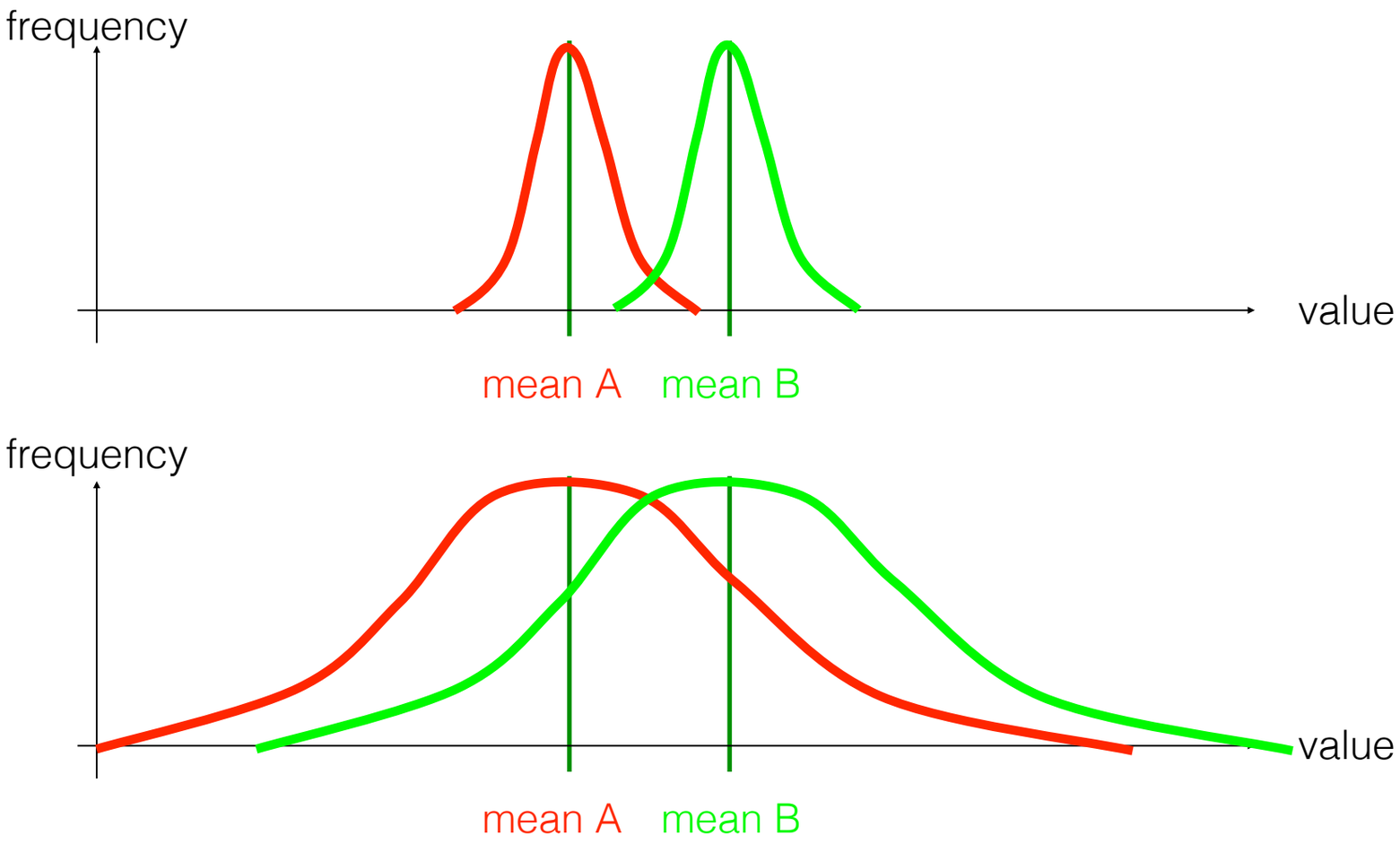
- Experimental conditions
  - 2 conditions – T9 and Key
  - Order of conditions is counterbalanced
    - User 1,3,5,7,9 perform T9 then Key
    - User 2,4,6,8,10 perform Key then T9
  - Different texts in first and second run?
  - Particular phone model?
  - Completion time is measured (e.g. stop watch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

Example: Study on Text Input

- 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
  - H_0-1: there is no difference in the input speed between keyboard and T9
  - H_0-2: there is no difference in the number of errors made using a keyboard input compared to T9

Comparing Values

Significant differences between measurements?



Significance

- In statistics, a result is called significant if it is unlikely to have occurred by chance.

- It does not mean that the result is of *practical significance*!
- A statistically significant speed difference of 0.1% between two text-entry methods may have little practical importance.

- In the case of hypothesis testing the significance level is the probability that the null hypothesis ('no correlation') will be rejected in error although it is true.

- Popular levels of significance are 5%, 1% and 0.1%

t-Test

- The t-test is a test of the null hypothesis that the means of two normally distributed populations are equal. The t-test gives the probability that both populations have the same mean (and thus their differences are due to random noise).
- A result of 0.05 from a t-test is a 5% chance for the same mean.
- Different variants of the t-test are used for paired (each sample in population A has a counterpart in population B) and unpaired samples.
- Examples:
  - Paired: speed of persons before and after treatment (within subjects design)
  - Unpaired: the reading speed of two different groups of people are compared (between subjects design)

Student [William Sealy Gosset] (March 1908). "The probable error of a mean". Biometrika 6 (1): 1–25.

Excel: t-Test

Real data from a user study

	A	B
K1	751	1097
K2	1007	971,5
K3	716	1121
K4	1066,5	1096,5
K5	871	932
K6	1256,5	926,5
K7	957	1111
K8	1327	1211,5
K9	1482	1062
K10	881	976
Mean	**1031,5**	**1050,5**
T-test	**0,8236863**	

	A	B
K1	826,5	1382
K2	806	1066
K3	791	1276,5
K4	896,5	1352
K5	696	1191
K6	1121	1066
K7	891	1217
K8	1327	1412
K9	1277	1266,5
K10	656	1101
Mean	**928,8**	**1233**
T-test	**0,0020363**	

Excel functions used:

=MITTELWERT(C4:C13)
=TTEST(C4:C13;D4:D13;2;1)

(function names are localized)
Menu: Tools>Data Analysis

TTEST(…) Parameters:

- Data row 1
- Data row 2
- Ends (1 or 2) (ususally 2)
- Type (1=paired, 2=same variance, 3=different variance)

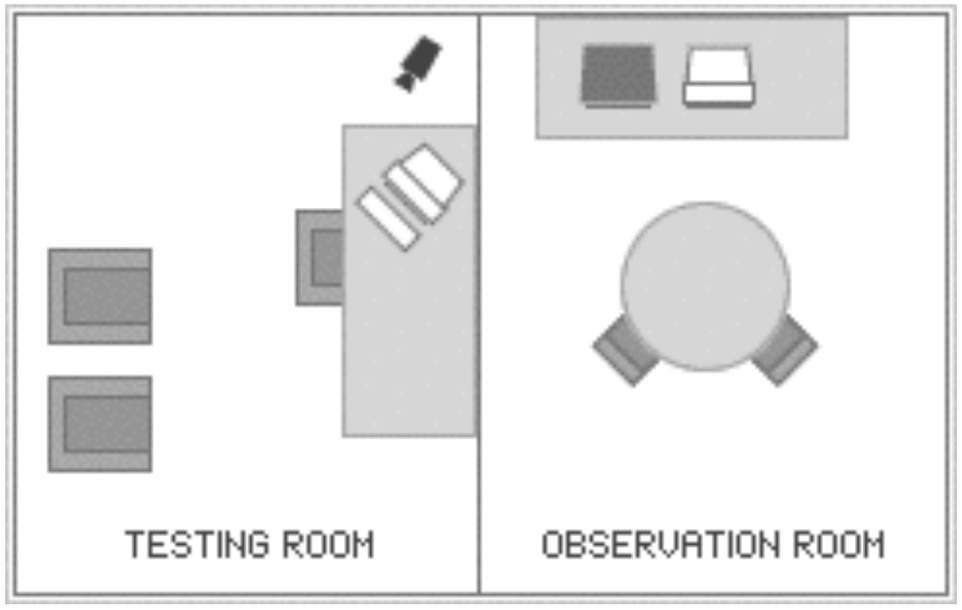
Analysis of Variance (ANOVA)

- Generalisation of the t-test
- Can cope with more than 2 data sets
- For 2 sets, basically the same as t-test => use t-test
- Can cope with more independent variables with multiple levels
- Multivariate ANOVA for more than one dependent variable
- Excel: http://office.microsoft.com/en-au/excel/HP100908421033.aspx

> “The experiment used a repeated measures within-participant factorial design 3 x 2 x 3 (interaction technique x transfer type x task type).”
>
> “The independent variable interaction technique consisted of three levels: standard Bluetooth, touch & connect and touch & select.”
>
> Khooviraj, Rukzio, Hardy, Holleis. MobileHCI’09

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



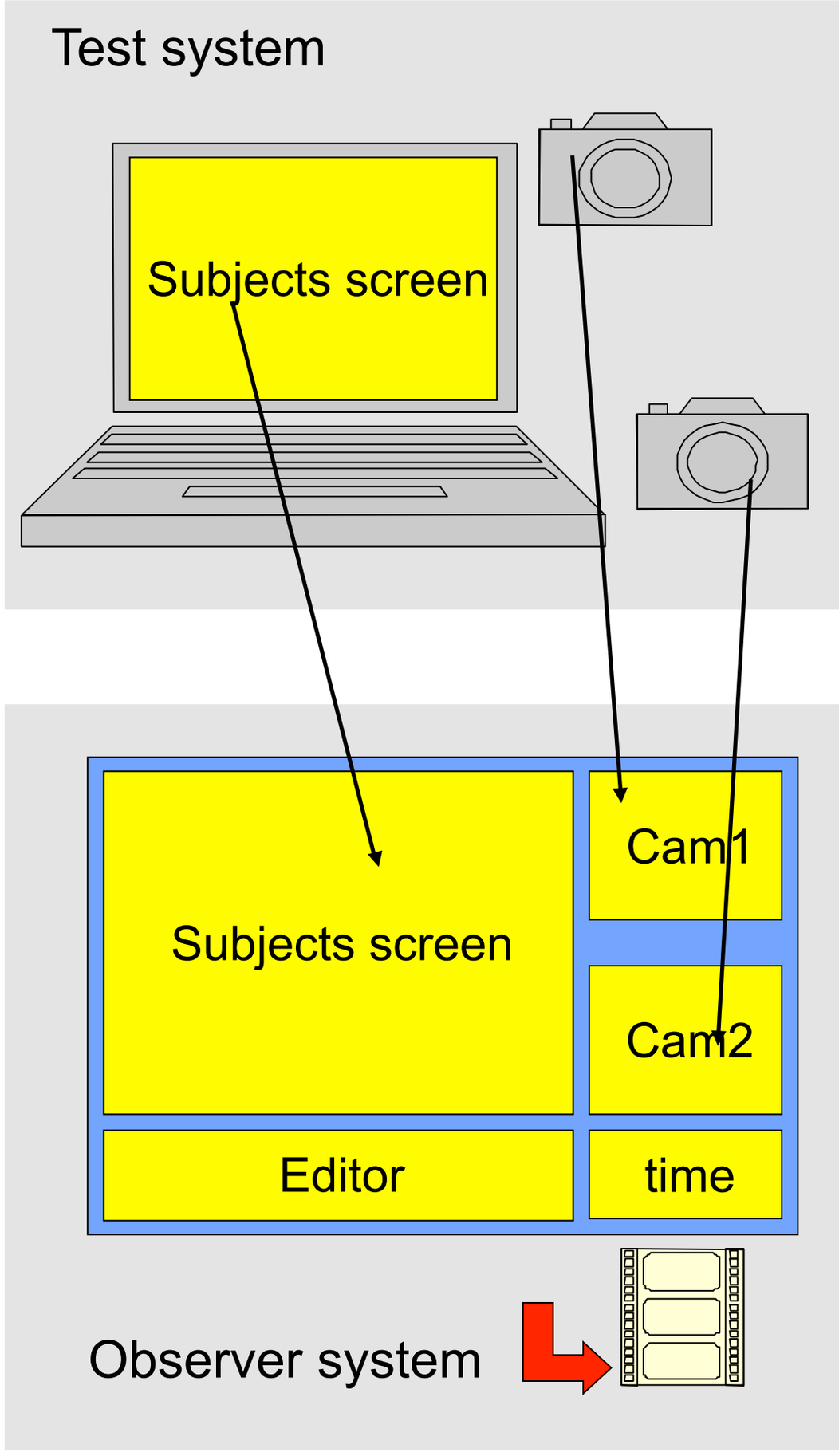
Source: www.xperienceconsulting.com

Poor Man‘s Usability Lab

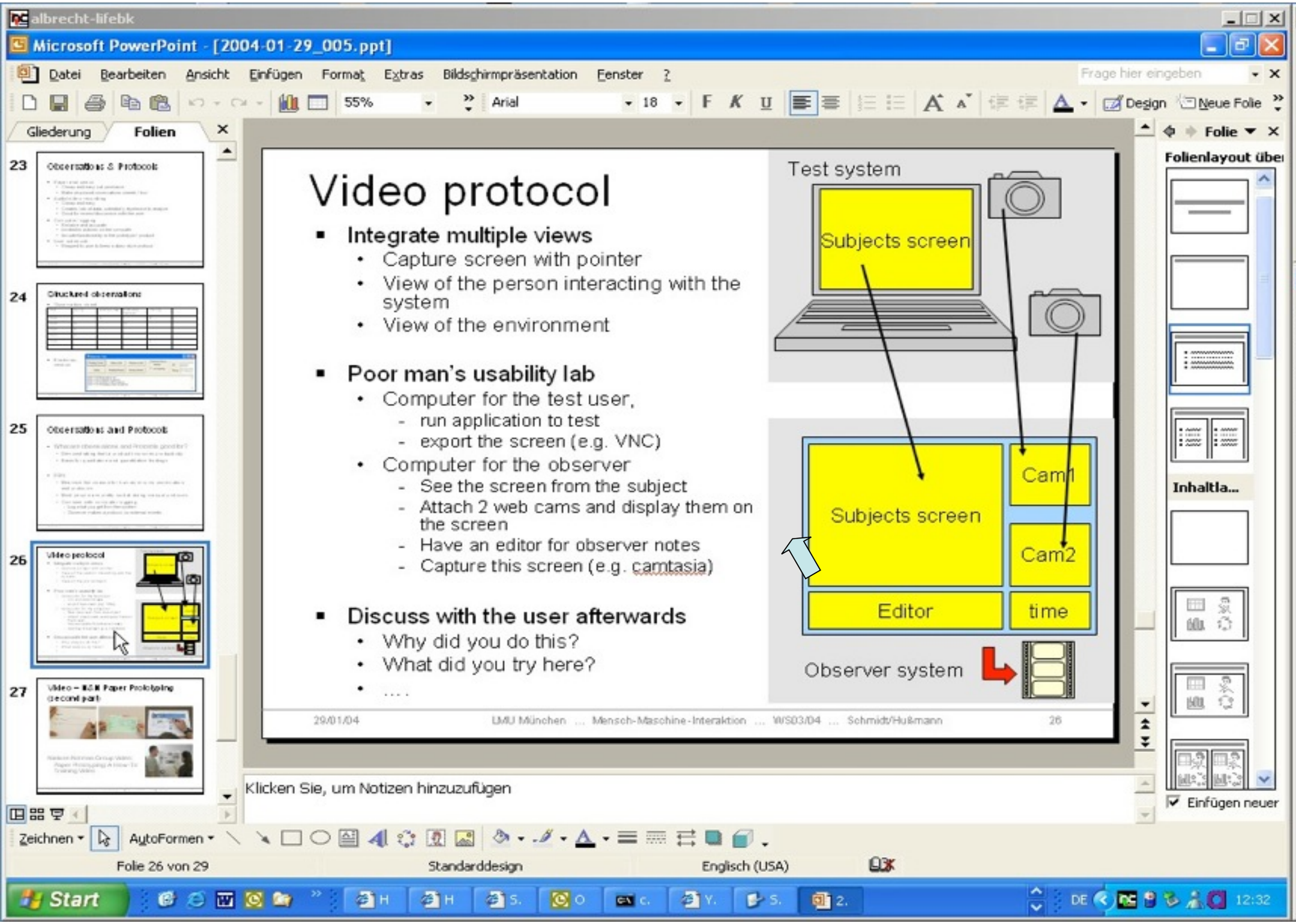
- Goal: Integrate multiple views
  - Capture screen with pointer
  - View of the person interacting with the system
  - View of the environment

- Setup:
  - Computer for the test user,
    - run application to test
    - export the screen (e.g., via VNC)
  - Computer for the observer
    - See the screen of the subject
    - Attach 2 web cams (face and entire user)
    - Display them on the observer‘s screen
    - Have an editor for the observer‘s notes
    - Capture this screen (e.g. QT, Camtasia)

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



Screen video



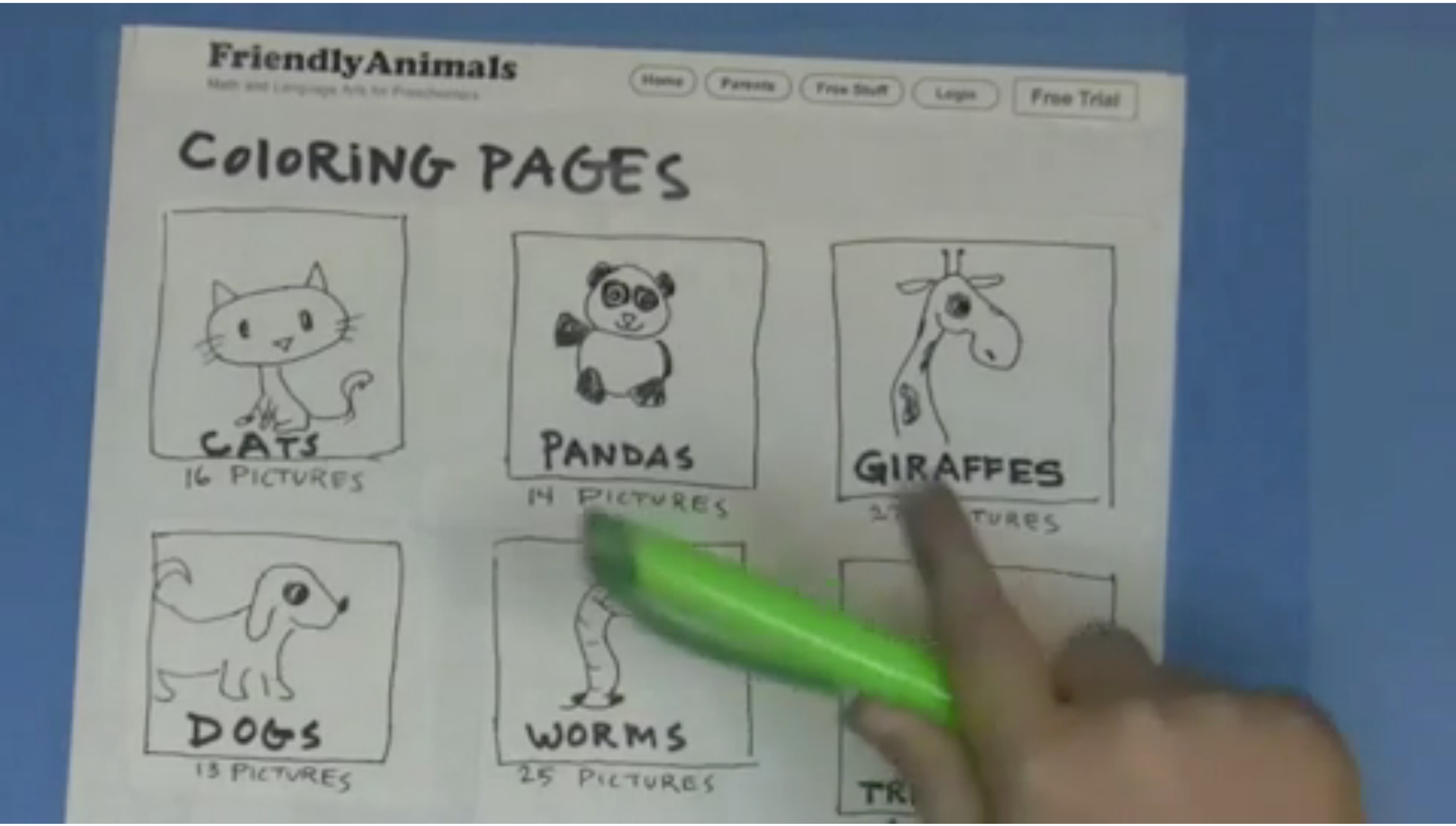
Field Studies

- 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

Agenda for this class

Intro & motivation		
	Formative	Summative
Analytical	Cognitive walkthrough	Heuristic evaluation
Empirical	Prototype user study	Controlled experiment Usability lab test Field studies
Discussion and take-home thoughts		

Paper Prototype Study



Agenda for this class

Intro & motivation		
	Formative	Summative
Analytical	Cognitive walkthrough	Heuristic evaluation
Empirical	Prototype user study	Controlled experiment Usability lab test Field studies
Discussion and take-home thoughts		

References

- 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/

- Further Literature
  - Andy Field & Graham Hole: How to design and report experiments, Sage
  - Jürgen Bortz: Statistik für Sozialwissenschaftler, Springer
  - Christel Weiß: Basiswissen Medizinische Statistik, Springer
  - Lothar Sachs, Jürgen Hedderich: Angewandte Statistik, Springer
  - various books by Edward R. Tufte

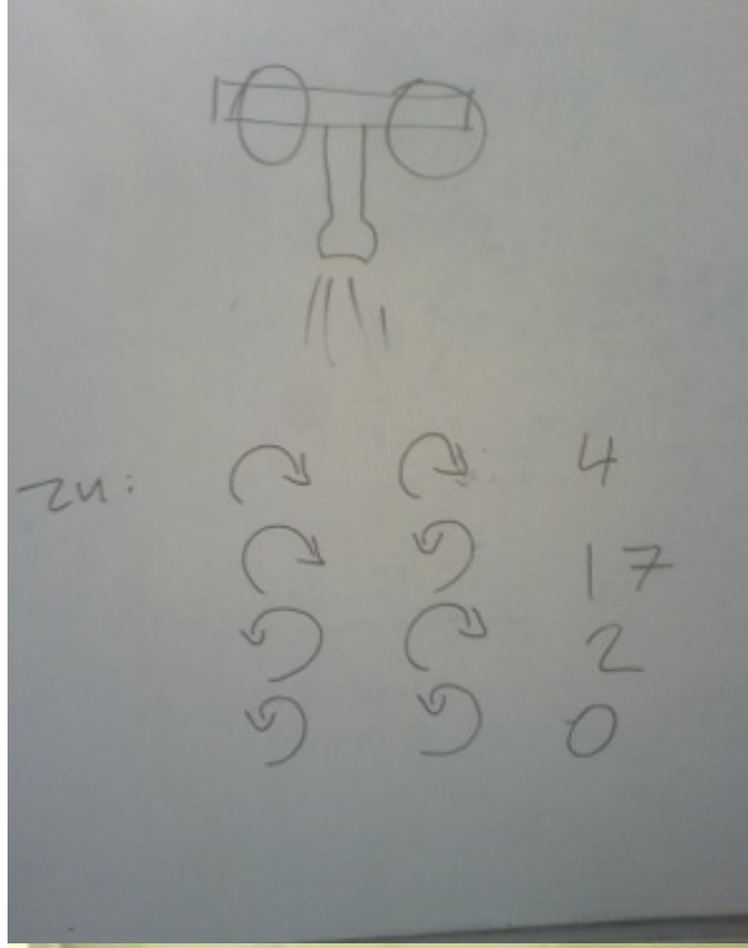
– video on next slide by Eric Shaffer, Human Factors Inc. :
http://www.youtube.com/watch?v=bminUlAu47Q


HOT


COLD

Intuitive Interfaces?

- Given: old style water faucet
  - 2 valves, 1 outlet
  - Cylindrical, next to each other
  - Left warm, right cold



- Question: In which direction does each valve close?

- Homework: find such faucets, determine which are „intuitive“ and why (not)