User Modeling 2007 – Tutorial 4

Formative Evaluation Methods for Adaptive Systems

Stephan Weibelzahl
Alexandros Paramythis
Judith Masthoff

Presenters

Dr. Stephan Weibelzahl
National College of Ireland Dublin
Mayor Street IFSC
Dublin 1, Ireland
+353 1 4498 579
sweibelzahl@ncirl.ie
http://www.weibelzahl.de/

Alexandros Paramythis
Institute for Information Processing and Microprocessor Technology (FIM)
Johannes Kepler University Linz
Altenbergerstr. 69
A-4040 Linz, AUSTRIA
+43 732 2468 8442
alpar at fim.uni-linz.ac.at
http://www.fim.uni-linz.ac.at/staff/paramythis/

Dr. Judith Masthoff
Department of Computing Science
University of Aberdeen
Aberdeen AB24 3UE
Scotland, UK
+44 1224 272299
jmasthoff at csd.abdn.ac.uk
http://www.csd.abdn.ac.uk/~jmasthof/

Time plan

13:30 - 14:15	**Introduction** **Layered Evaluation**
14:15 - 14:45	**1st Hands-on Session**
14:45 - 15:00	**Evaluation Methods** (Part A)
15:00 - 15:30	*Coffee break*
15:30 - 16:15	**Evaluation Methods** (Part B)
16:15 - 16:45	**2nd Hands-on Session**
16:45 - 17:00	**Common Pitfalls** **Where to next?**

Introduction

Evaluation in general

- Evaluation is the systematic determination of merit, worth, and significance of something or someone

- In this context: all types of user studies that
  - » inform the development or improvement of a system
  - » demonstrate the impact of a technology

- Examples include
  - » Experiments, Case-studies, Surveys, Usability studies, Expert ratings, etc.

Why is evaluation important?

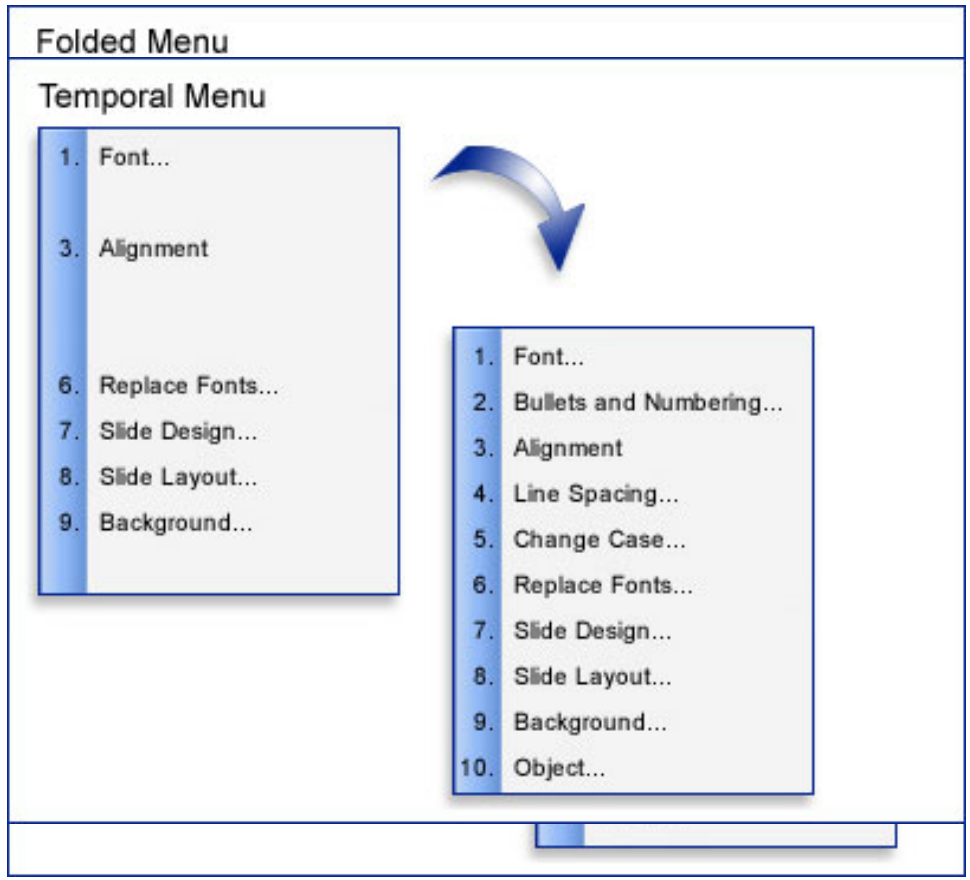
- Find out whether it really works
  - » Effectiveness
  - » Efficiency
  - » Usability, user satisfaction
- Detect inaccuracies and invalid assumptions
- Convince users, customers, investors, PhD examiners
- Scientific advancement

Why is the evaluation of adaptation different?

- Basic premise of “traditional” HCI evaluation:
  - » All users experience the *same* system
- Basic premise of interactive adaptive systems:
  - » Each user experiences a *personalised* version of the system
- Approaches that have been tried and have only been partially successful:
  - » “With and without” adaptivity
  - » Adaptivity as a single system feature

An (elusively) simple example

- Evaluate the employment of adaptive menus in a word processing application
  - » Adapt how?
  - » Adapt when?
  - » Based on what?
  - » Level of user control?
  - » Why do some users love it and others hate it?



Images from: Kath Straub (2004) Adaptive Menu Design, UI Design Newsletter – July, 2004

Tutorial goals

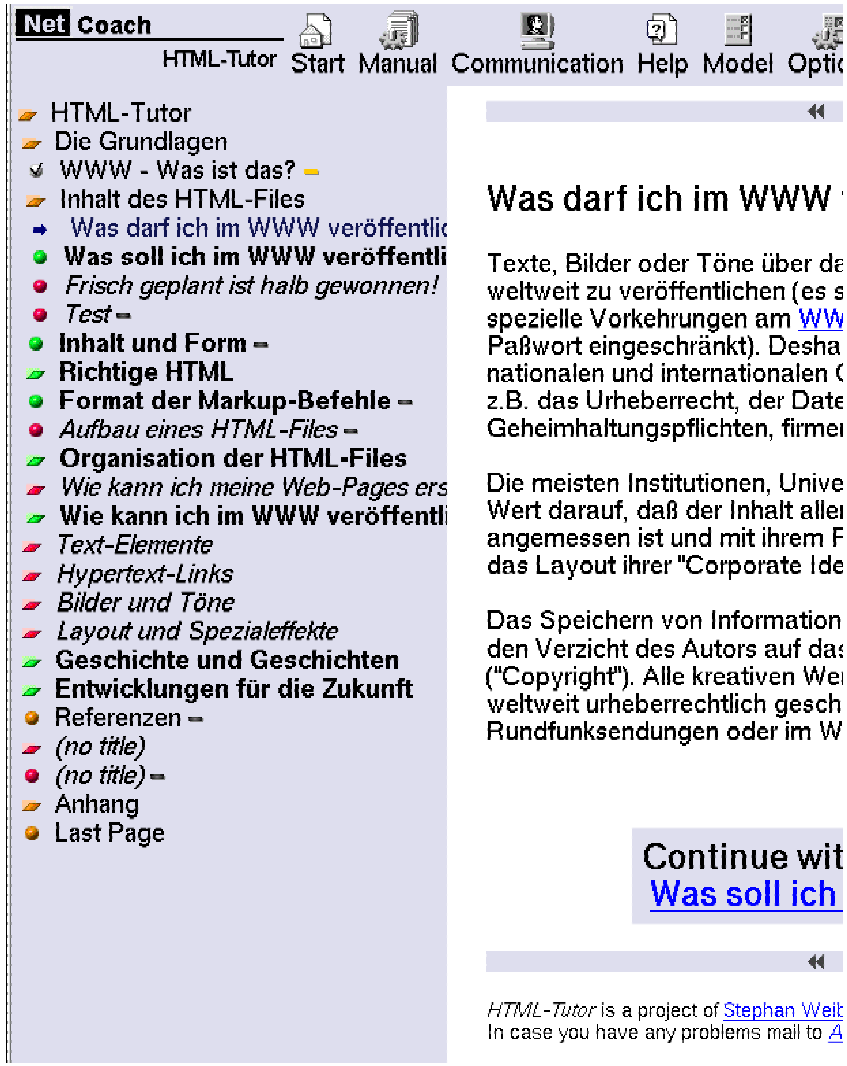
- Upon successful completion of the tutorial, participants will
  - » Be aware of the specific problems involved in the evaluation of adaptive systems that differentiate them from their non-adaptive counterparts, and able to solve or circumvent these problems
  - » Understand and be able to apply the principles of layered evaluation of adaptive systems
  - » Be able to design a targeted formative evaluation study for an adaptive system (e.g., addressing a given layer and set of criteria) by selecting appropriate methods and criteria

Introduction of Layers

Identifying evaluation layers and criteria

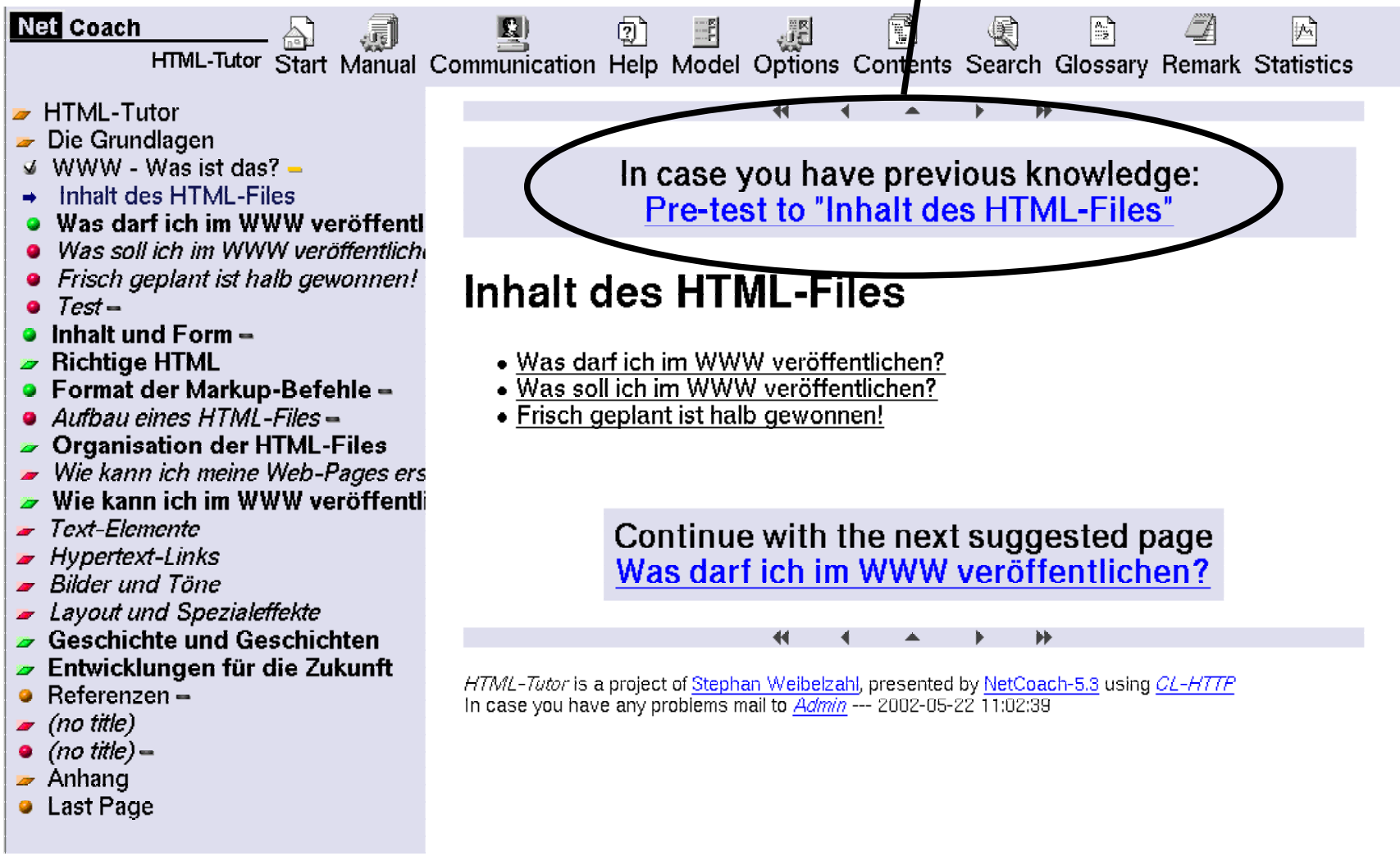
Example Study

- HTML-Tutor: An Adaptive Learning System
- Introduction to HTML and Publishing on the Web

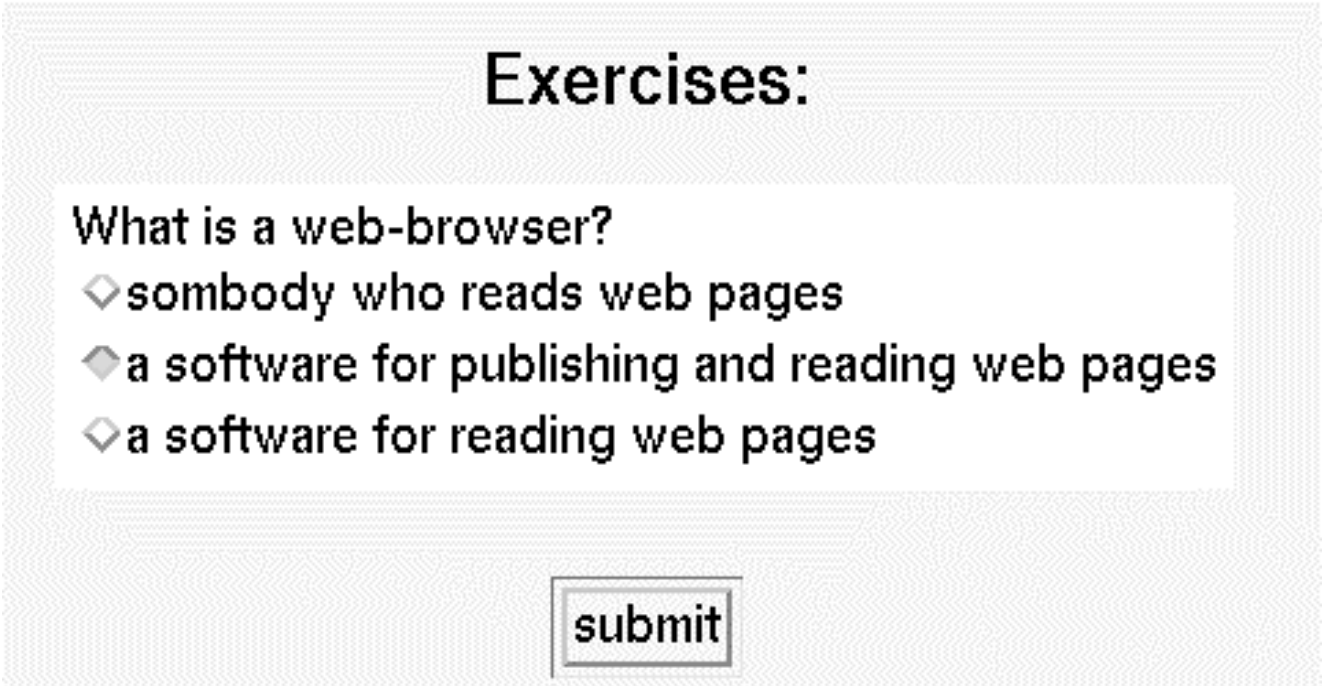


Pre-Tests

optional pretest



Exercises


Exercises:
What is a web-browser?
sombody who reads web pages
a software for publishing and reading web pages
a software for reading web pages
submit

Exercise Feedback

The question was:

What is a web-browser?

Your answer was wrong:
A: Your answer
S: correct solution

A	S	
◇	◆	a software for reading web pages
◆	◇	a software for publishing and reading web pages
◇	◇	sombody who reads web pages

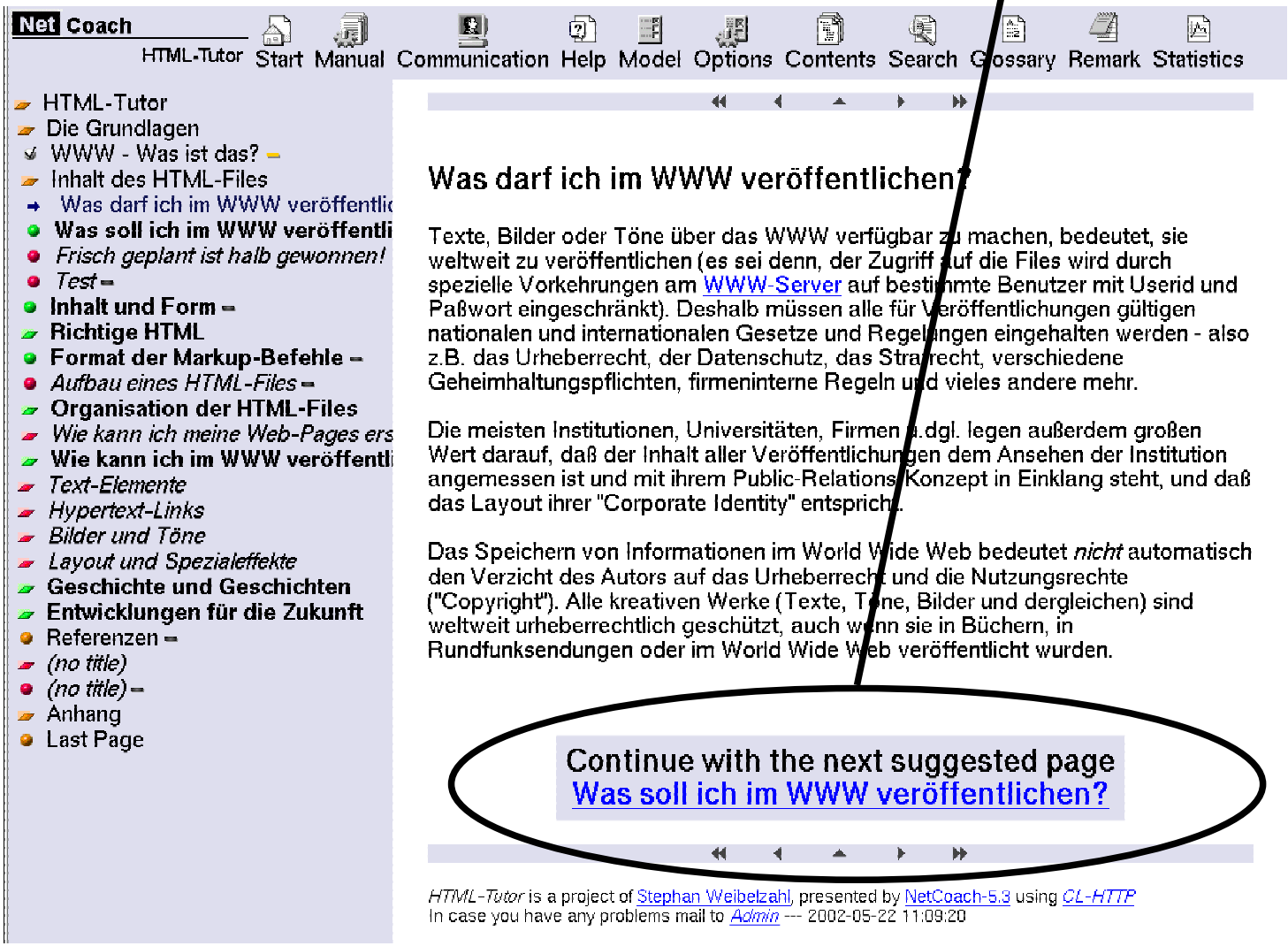
Reason:
A web browser is a software for reading web pages. To publish a web page a web server is required.

Adaptation Strategy

adaptive link annotation

Adaptation Strategy

adaptive curriculum sequencing



How can I evaluate my system?

- How to evaluate HTML-Tutor?
- How can we find out what's wrong?
- How to improve it?

- Compare adaptive version to non-adaptive version of the course?
  - » What could we learn from that?
  - » What can we not learn from that?

Layered evaluation

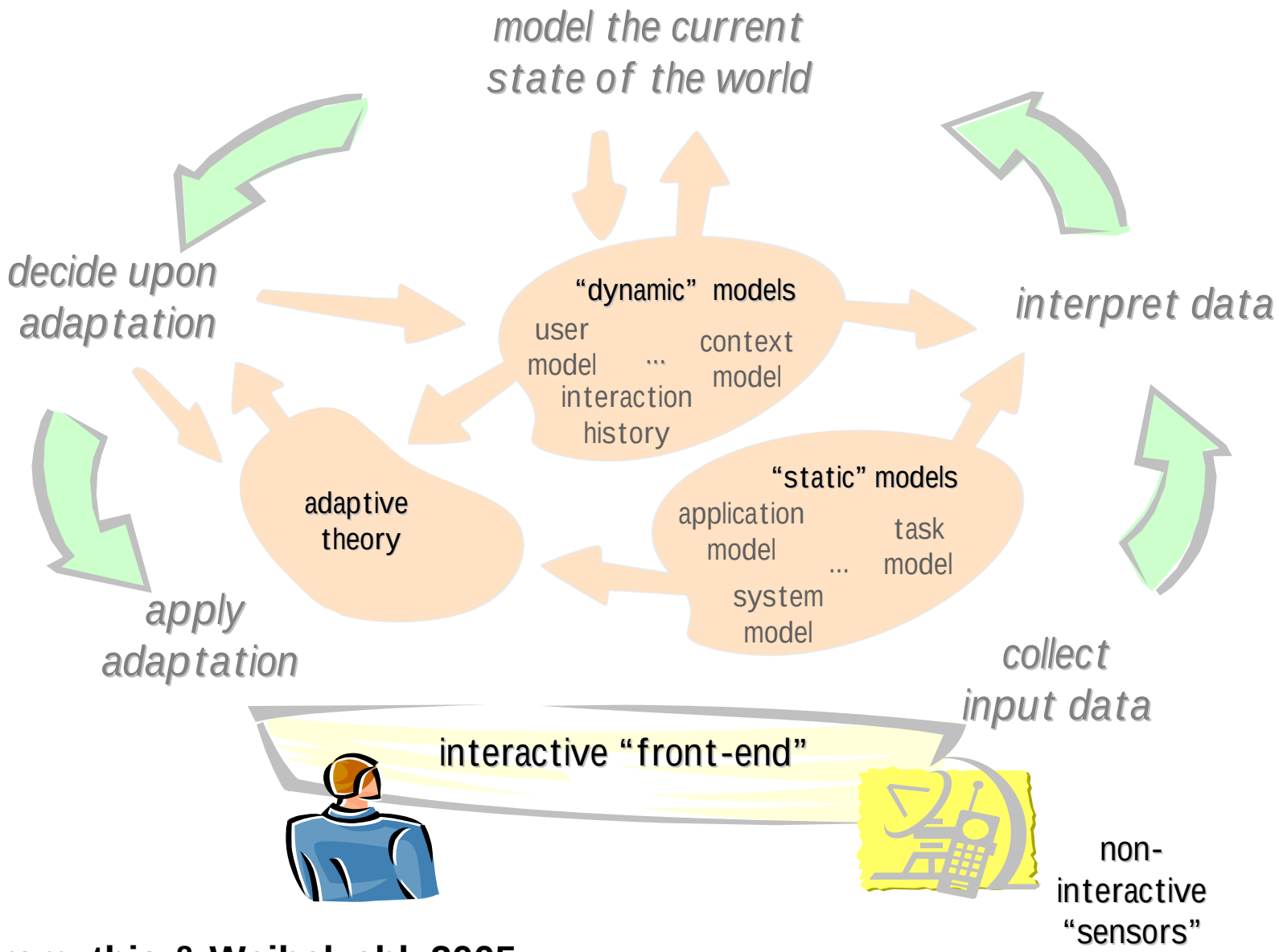
- Basic premises of layered evaluation
  - » Don’t treat adaptation as a “monolithic” / singular process (at least not only as such)
  - » Rather, “break it down” into its constituents (“layers”), and
  - » Evaluate each of them separately where necessary and feasible

- Basis for this tutorial:
  - » Paramythis and Weibelzahl, 2005 (“the merger”)
  - » Origins
    - Paramythis, Totter, and Stephanidis, 2001
    - Weibelzahl and Weber, 2001

How can I evaluate my system?

- Important decisions
  - **Evaluation layer(s): What to evaluate?**
  - **Criteria: What are measures of success?**
  - Method: How to collect data? (see “Methods” part)

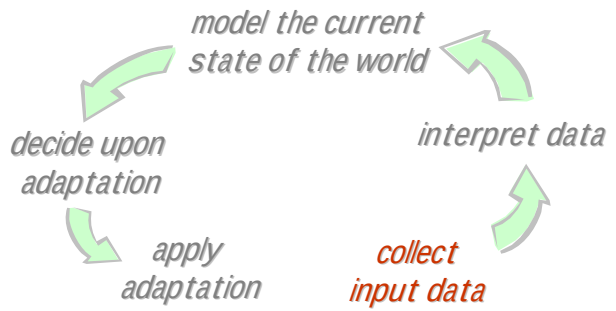
Layered Evaluation



Paramythis & Weibelzahl, 2005

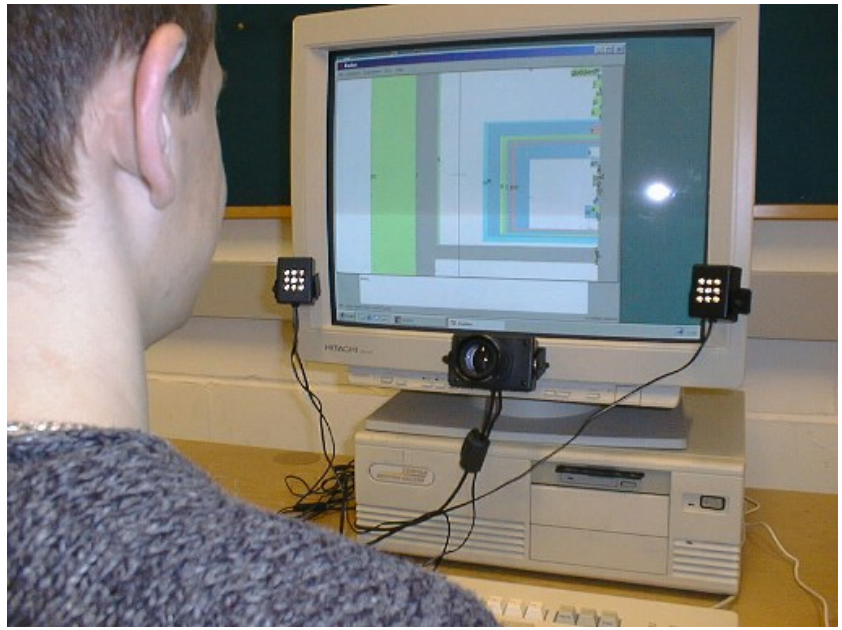
Collection of Input Data

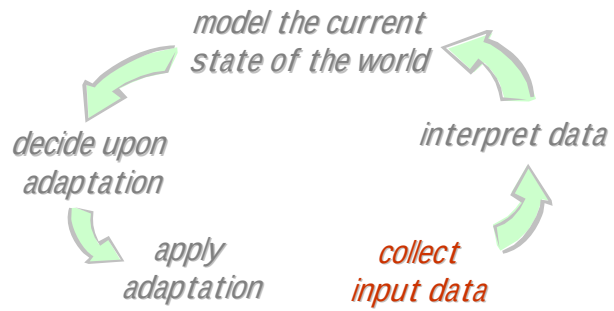
- Adaptive system observes user behaviour and context, e.g., click stream, input, sensor data, etc.
- Questions
  - » Does the data collection work?
  - » Is the user behaviour registered accurately?
- Criteria
  - » Reliability (consistency of data)
  - » Accuracy
  - » Latency



Collection of Input Data

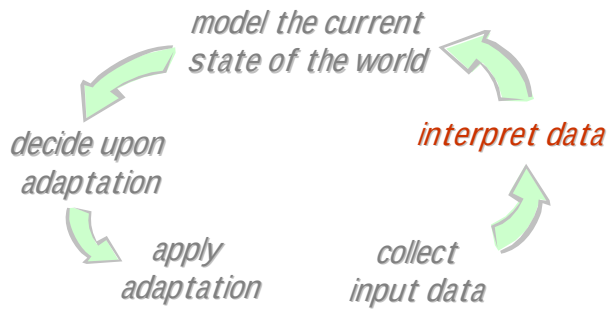
- Examples of questions to be answered
  - » Eye-tracking for task detection: Does the user actually look at the part of the screen that the eye-tracker indicates?
  - » HTML-Tutor: Are test items reliable?
  - » Movie recommender: Are ratings of movies consistent per user? Would a user rate the movie in the same way again after one week?





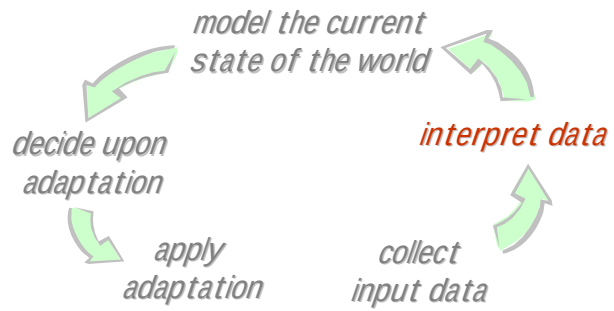
Interpretation of the Collected Data

- Adaptive system interprets the recorded behaviour
- Giving meaning to raw data
  - » Sometimes trivial (click on “next” button means, user wants to proceed to next page)
  - » However, interpretation is possibly based on assumptions and might require inference
- Question
  - » Are the users doing what the system thinks they are doing?
- Criterion
  - » Validity



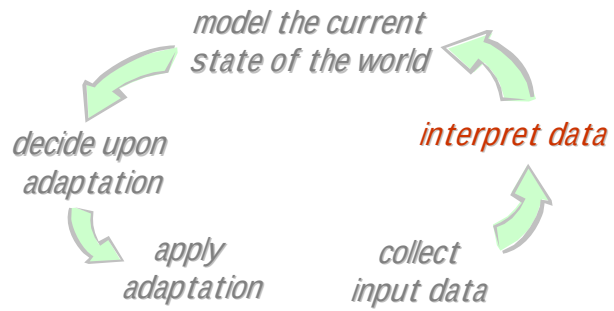
Interpretation of the Collected Data

- Examples of questions to be answered
  - » HTML-Tutor: Is the content of a page actually “known” when
    - Learner visited the page
    - Learner answered test- items correctly
  - » Movie recommender: Does a user actually like a movie when giving a positive rating?



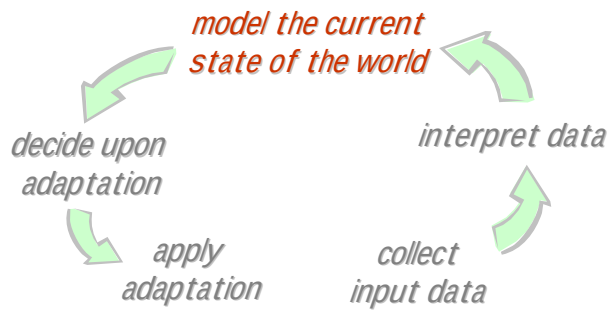
Example Study with HTML-Tutor

- Learners use system
- Learners complete post-test
- Comparison of model (“visited”, “known”) and real data



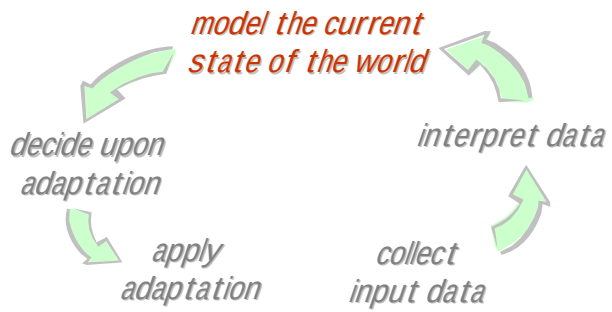
Modelling of the Current State of the World

- Based on observations the system infers the current state of the world, e.g., user model, context model
- Usually this is the AI component of the system (Bayesian network, rules, etc)
- Questions
  - » Does the model reflect the real world?
  - » Is the world modelled in an appropriate way?
- Criteria
  - » Primarily: Validity
  - » Secondary: Comprehensiveness, Redundancy, Precision, Sensitivity, Scrutability



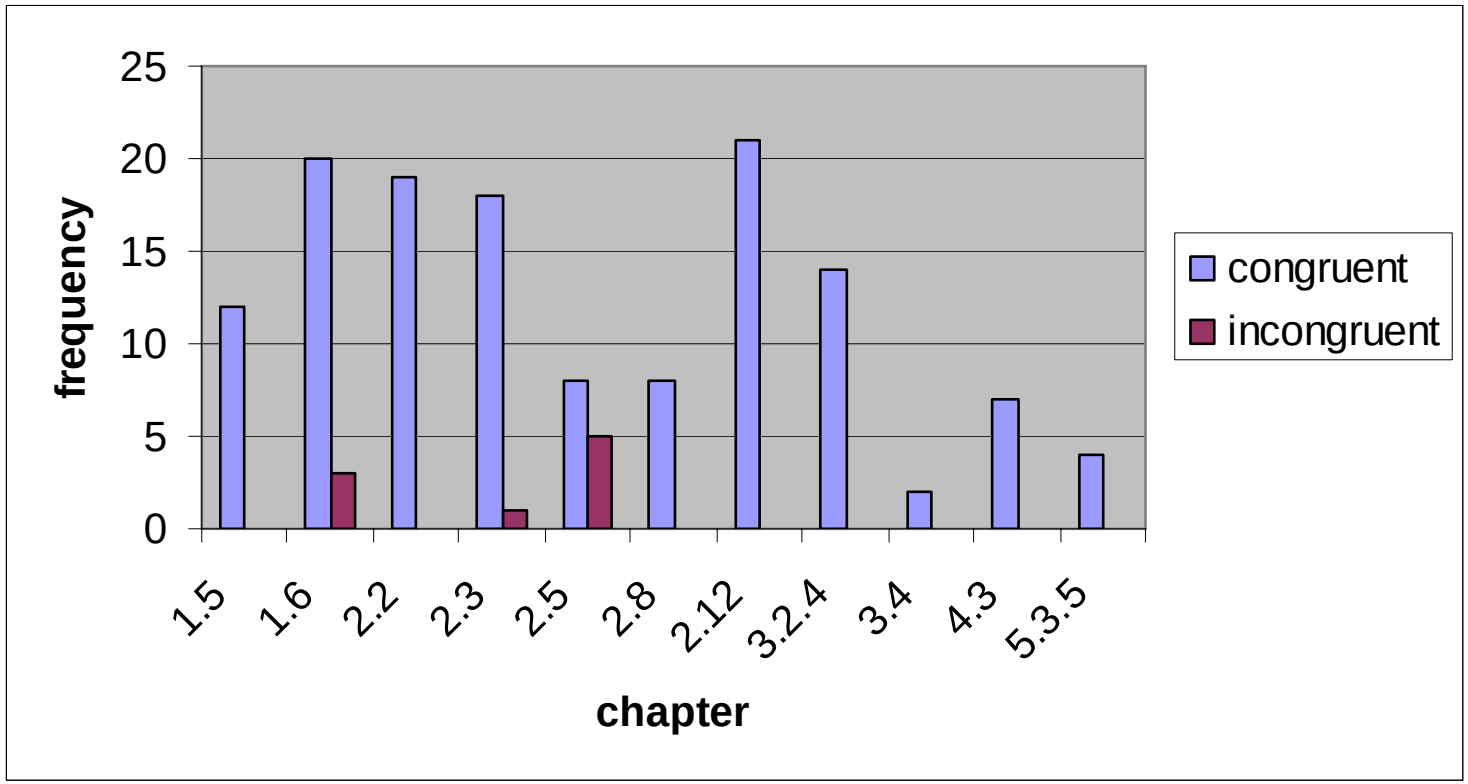
Modelling of the Current State of the World

- Examples of questions to be answered
  - » HTML-Tutor: Are pages that are inferred to be “known” (e.g., prerequisites of more advanced concepts) actually known?
  - » Movie Recommender: Do users like a movie that got high ratings from somebody with similar preferences?



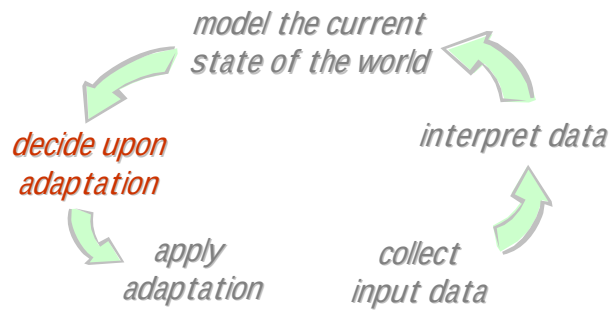
Example Study with HTML-Tutor

- Learners learn concepts in class
- Learners use system
- Learners complete post-test
- Comparison of model (“inferred”) and real data



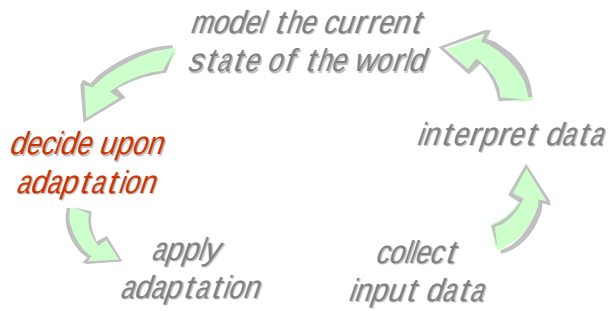
Decide upon Adaptation

- Adaptive System decides which adaptation theory/strategy to apply given the current user model
- Questions
  - » Is it necessary to intervene?
  - » Did the system select a good and appropriate adaptation strategy?
- Criteria
  - » Necessity
  - » Appropriateness
  - » Subjective acceptance



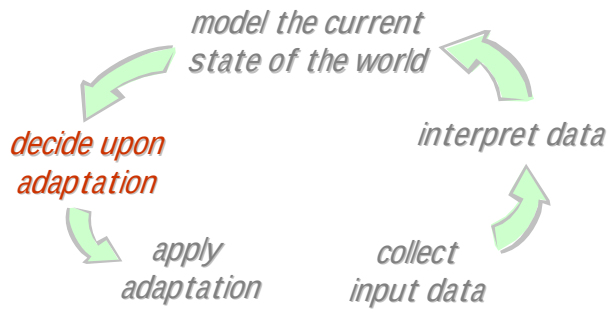
Decide upon Adaptation

- Examples of questions to be answered
  - » HTML-Tutor: The learner model seems to indicate that the learner acquired sufficient knowledge about the current chapter.
    - Shall we recommend to proceed to the next chapter?
    - Shall we annotate the current chapter as “known”?
  - » Movie recommender: Shall we recommend a certain movie (push) or wait till the user asks for a recommendation (pull)?



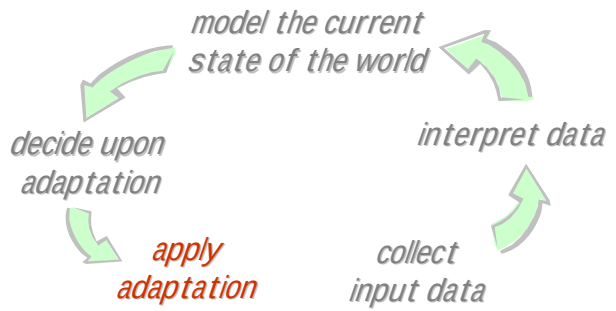
Example Study

- Learners use system under different conditions
  - » With and without annotation
  - » With and without sequencing
- Results
  - » No effect on number of pages visited, overall impression or perceived successful adaptation
  - » Annotation increases number of pages visited per minute
- How could this study be improved to better fit the layer?



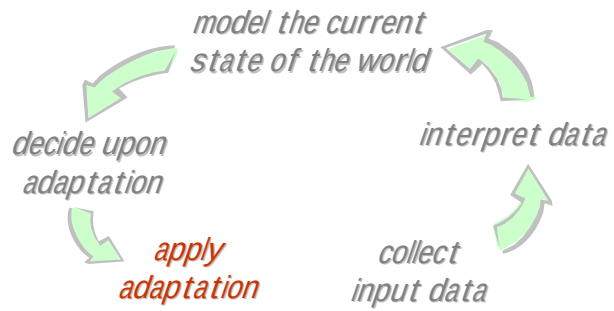
Applying Adaptation Decisions

- The adaptation decision can be applied in different ways (e.g., different colours, layouts, formulations)
- Questions
  - » Is the concrete instantiation of the adaptation decision working?
  - » Do users understand what it means?
  - » Do they like it?
- Criteria
  - » Usability
  - » Obtrusiveness
  - » Acceptance
  - » Timeliness
  - » User control



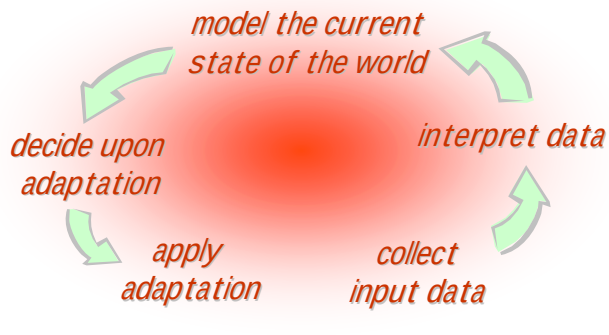
Applying Adaptation Decisions

- Examples of questions to be answered
  - » HTML-Tutor:
    - Is a red bullet a good way to indicate a “not recommended” page?
    - “Continue with the next suggested page”?
  - » Movie Recommender:
    - Shall we provide the full list of recommended movies?
    - Only one movie plus “more” button?
    - “Based on your ratings we believe that you might like the following movies…”?



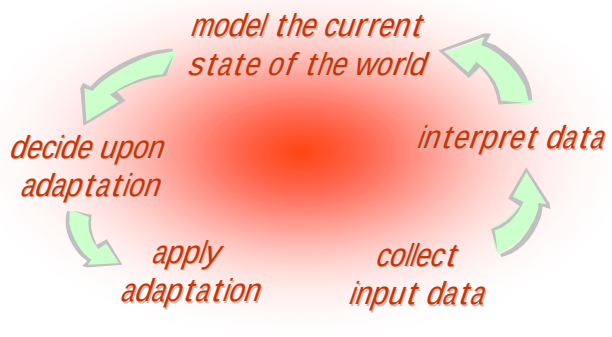
Evaluating Adaptation as a Whole

- The big picture
- Looking at the system as a whole: Does it work?
- Questions
  - Does the system achieve its goals?
  - Does it improve interaction?
  - Do users like the system?
- Criteria
  - Effectiveness
  - Efficiency
  - Usability
  - **System specific criteria**



Evaluating Adaptation as a Whole

- Examples of questions to be answered
  - » HTML-Tutor: Does adaptation to prior-knowledge save time?
  - » Movie Recommender: Do users find movies they like and would they have found these movies otherwise?



Field Study

- What's the impact of offering an adaptive prior-knowledge test in an on-line course? (Weibelzahl & Weber, 2002)
- 140 users learned with the HTML-Tutor
  - » optional pre-test for 3 chapters
  - » final knowledge test at the end of the course
  - » criteria: duration, knowledge
  - » statistical analysis: MANOVA and ANOVA

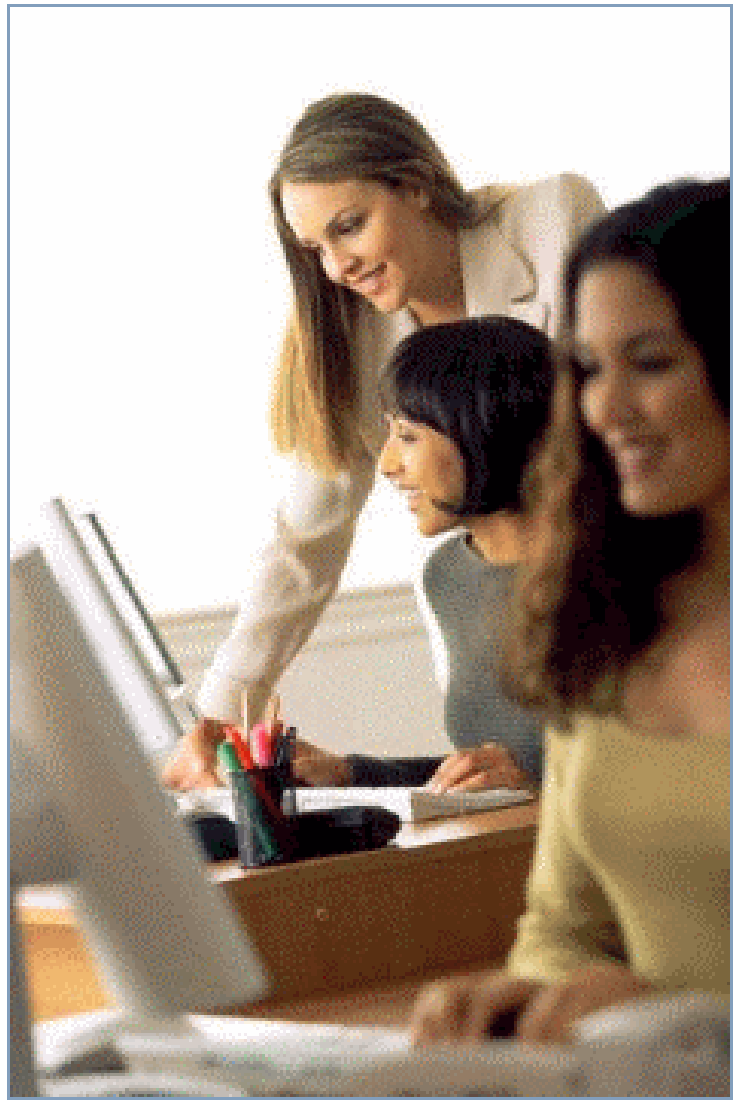
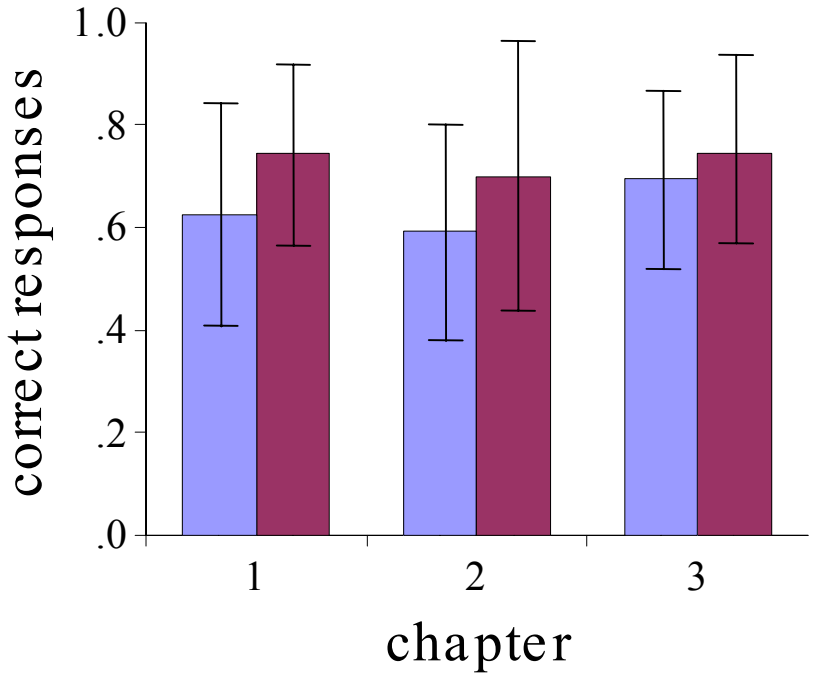
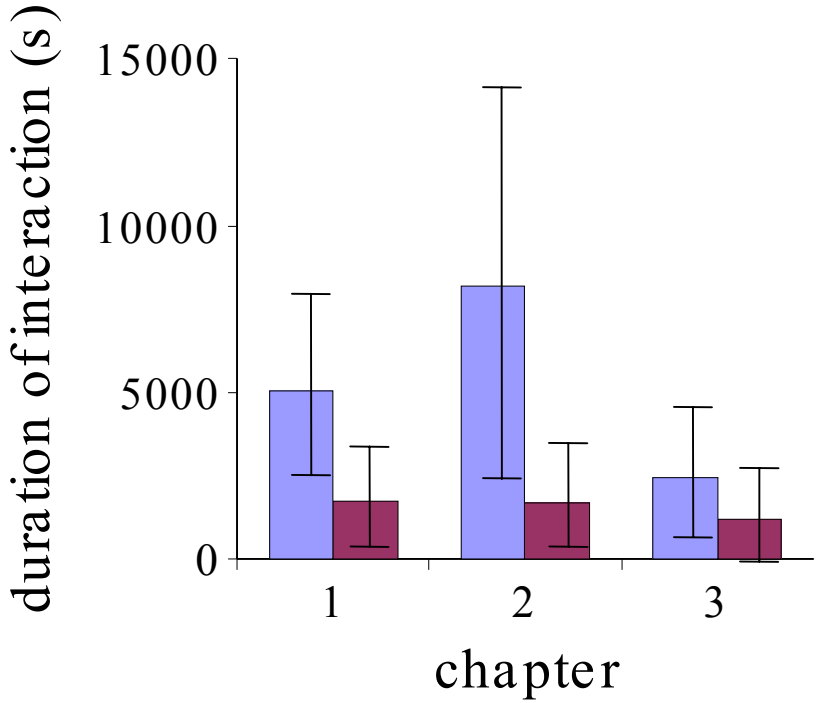


Photo © BrowserBob, 2007

Results

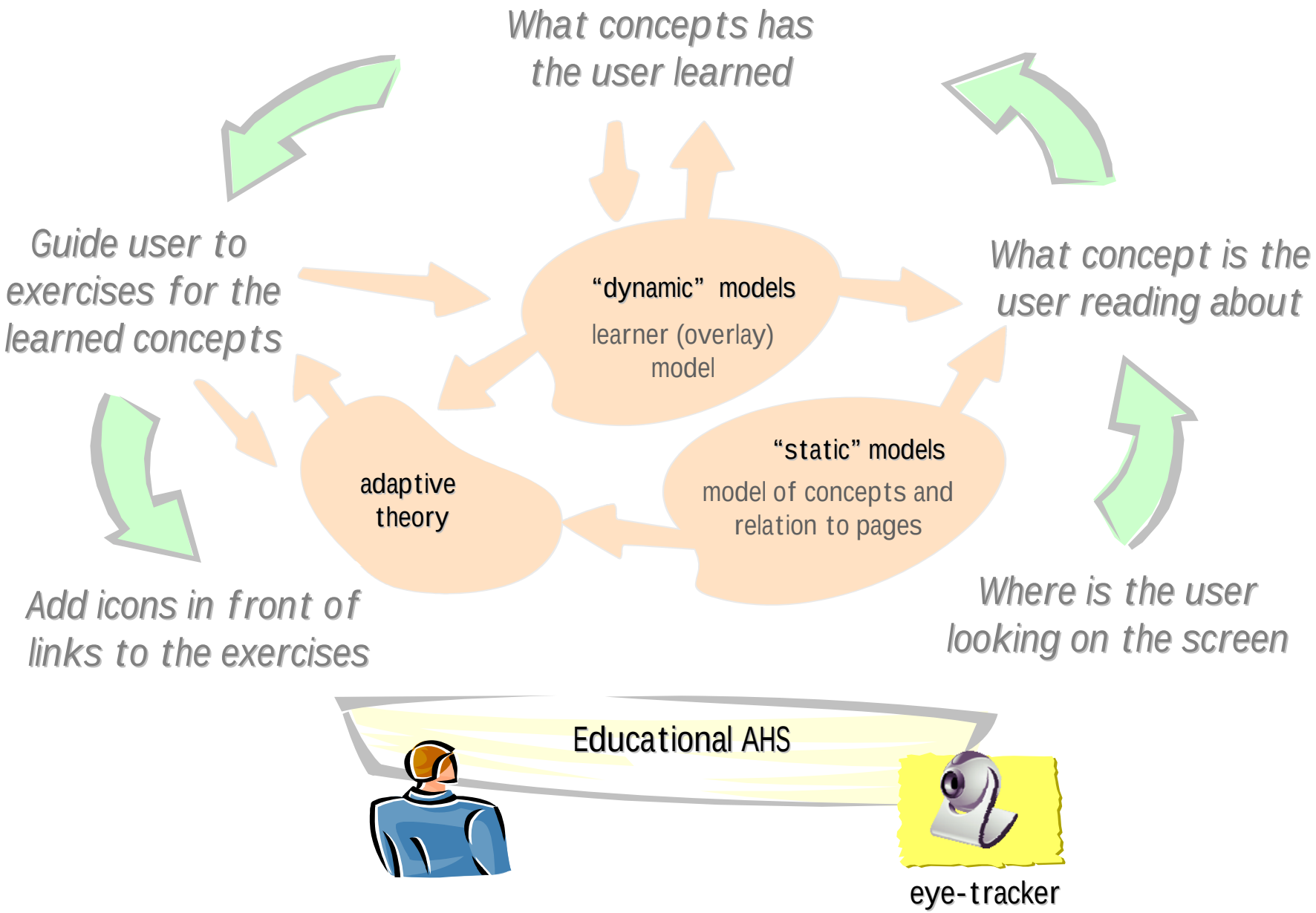
- No differences in knowledge
- Completed course much quicker





no pre-test presented
pre-test presented

Layered Evaluation – Recap



Layered Evaluation Summary

- Break adaptation process down into its constituents (“layers”)
- Evaluate each of them separately where necessary and feasible
- It’s meant to provide guidance rather than prescribing a certain way of evaluation
- Benefits
  - » Offers guidance for possible studies (“separation of concerns”)
  - » Helps to identify problems and wrong assumptions
  - » Guides development process (formative evaluation)

“Hands-on” Session

1st Part

Identifying evaluation layers and criteria

“Hands-on” session overview – 1st Part

- Goal
  - Apply what you have learned in typical adaptive systems
- Two parts
  - 1st part – Identify layers and establish evaluation criteria
  - 2nd part – Select evaluation methods and data collection instruments
  - The output of the first session will be used as input for the second
- Organisation of first part
  - Brief presentation of the systems
    - Adaptive super market
    - Adaptive music player suite
    - Adaptive email classification system

"Hands-on" session overview – 1st Part

- Organisation of 1st part *(cont)*
  - System "leaflets"
  - Selection of a system to work with
  - Separation into groups
  - Group discussion / work
  - Sampling / presentation of results

- Expected output
  - Description of how layers apply to your system
  - A list of evaluation criteria for each layer to be addressed
  - A list of domain-specific evaluation topics and criteria for the system as a whole

- Time available: 30 minutes

Target evaluation systems

- **The Centaur adaptive super market**
  - **Main adaptive functions**
    Personalised product recommendation
  - **Monitored data**
    User's browsing, searching, purchasing behaviour
  - **Behind the scenes**
    Classification learning, collaborative filtering

Photo © Spyros Vagelakis, 2004

Target evaluation systems (cont)

- **The Ananas adaptive music player suite**

  - **Main adaptive functions**
    Recommend music that fits the user's musical tastes, affective state, and context

  - **Monitored data**
    Music metadata, listening patterns, playlists, physiological indicators

  - **Behind the scenes**
    Decision theoretic approach, collaborative filtering, and rules

Target evaluation systems (cont)

- **The Trippy adaptive email classification system**

  - **Main adaptive functions**
    Automatically determine the folder in which a user would place a given email, and facilitate the process of actually placing it there

  - **Monitored data**
    Emails already in folders, and the user's response to the system's recommendations

  - **Behind the scenes**
    Classification learning, possibly in combination with utility functions

Questions to focus on

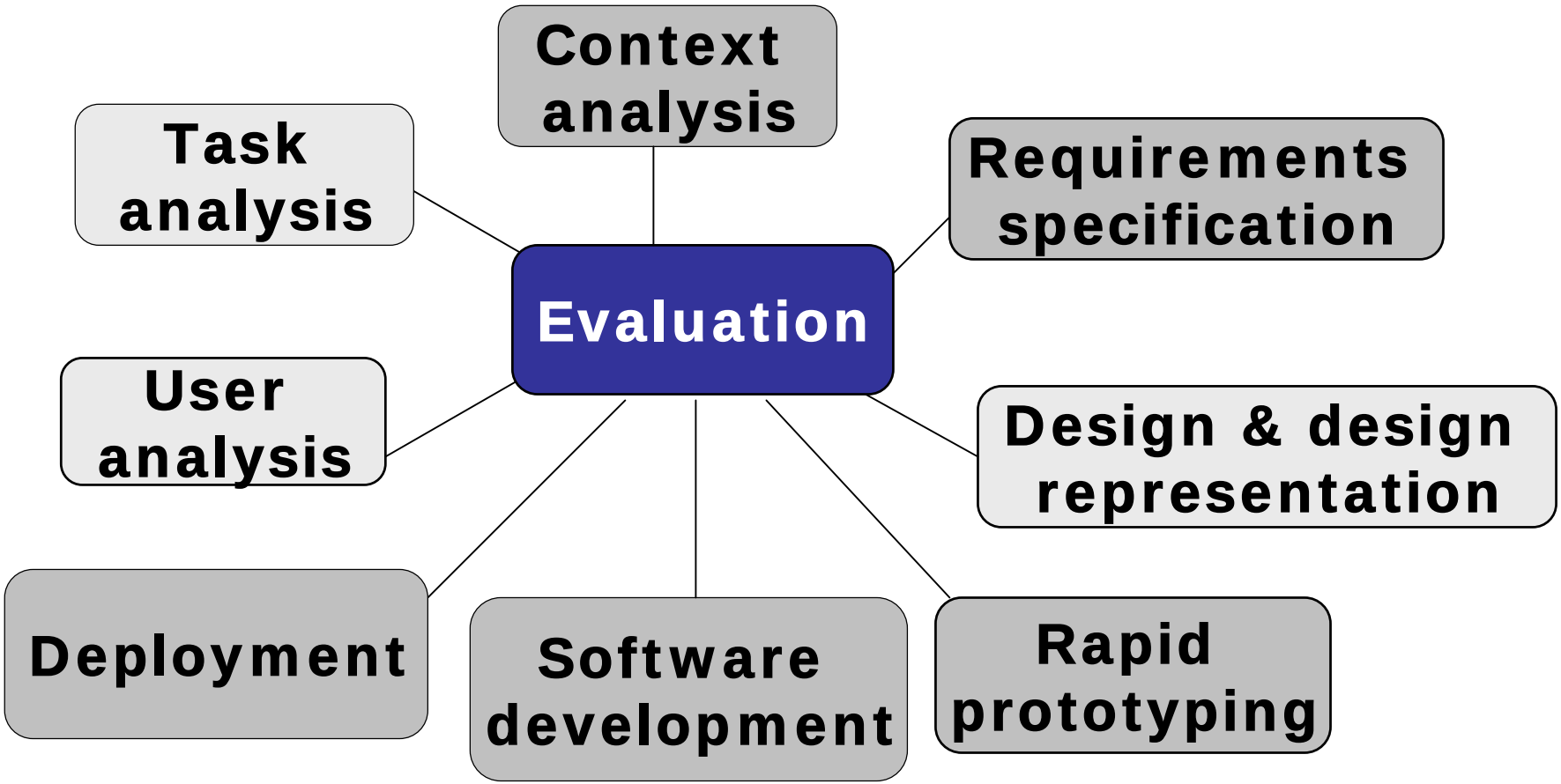
- Overall
  - » What needs to be evaluated in the system, as far as adaptivity is concerned?
  - » We are **not** concerned yet with the how
- Layer-specific
  - » Which layers are implicated in each of the above cases?
  - » What are the relevant criteria for each of the layers?
- Domain-specific
  - » What criteria can be used to judge whether adaptivity meats it’s goals in the context of the specific system (or in the system’s domain more generally)?

Evaluation Methods

Different methods

- When the evaluation is done
- How the evaluation is done
- By whom the evaluation is done

When the evaluation is done



Evaluation throughout!

When the evaluation is done (cont)

- Different methods applicable depending on the stage of development

- We will distinguish two main points for evaluation
  - » Design (GUI and/or algorithms) has been done
  - » Prototype has been implemented

- However, two of the methods (Focus Group and User as Wizard) can be applied even earlier, to inspire algorithm

How the evaluation is done

To evaluate a layer, you need to know:

- What input it receives
  - » Show the evaluator what the input is, OR
  - » Let the evaluator decide the input

  Could be input over long period of time

- What output it produces
  - » May require effort, as most layers will not have a GUI
  - » Outputs may be hard to distinguish: difficult to look at Apply Adaptation separate from Decide Adaptation

How the evaluation is done (cont)

Different ways of evaluating:

- Analyse strengths and weaknesses
- Compare against criteria
- Perform tasks

By whom the evaluation is done

- Users
  - » Most realistic, as they will end up using the system
- Experts
  - » May be needed, because too difficult for user (e.g. if input/output is a Bayesian net)
  - » May understand criteria better
- Simulated Users

Evaluation Methods Overview

Layers' Input	Layers' Output	Type of Evaluator	Measuring Method	Evaluation Method
Shown	Shown	Users or Experts	Discussion	Focus Group 
Shown	Shown	Experts	Criteria 	Cognitive Walk. Heuristic Eval. 
Decided / Shown	Shown	Users or Simulated users	Task performance	Task-based experiment  Simulated users 
Decided	Shown	User or Experts	Interview or Criteria	Play with layer 
Shown	Decided	Users or Experts	Comparison with system or Criteria	User/Expert as Wizard 

Evaluation Methods

- Focus Group
- Cognitive Walkthrough
- Heuristic Evaluation
- Task-based Experiment
- Play with Layer
- User/Expert as Wizard
- Simulated Users

Focus Group – Traditional

- Show a group of users a prototype and ask their opinion

Focus Group on layer's performance

- Discussion of a layer's performance in the informal setting of a focus group
- Show the input to the layer, and the output of the layer, in a way understandable to participants
- As adaptation is dynamic (and taking time), the input may have been received over a prolonged period (e.g. sequence of events)
- Depending on the implementation, participants may have to be experts (e.g. if UM is a Bayesian net)

Example: UM Layer of ITS

- Input: Interpreted User actions
  - The user got 18 out of 20 items right on a test of IF-THEN statements.
  - Next, the user got 3 out of 20 right on a test of WHILE statements.
  - Finally, the user got 2 out of 20 right on a test of assignment statements

- Output: User Model
  - Mastery of IF-THEN statements 9, confidence 9
  - Mastery of WHILE statements 2, confidence 9
  - Mastery of assignment statements 1, confidence 9
  - Emotional state: demotivated, confidence 6

Example: DA Layer of Recommender

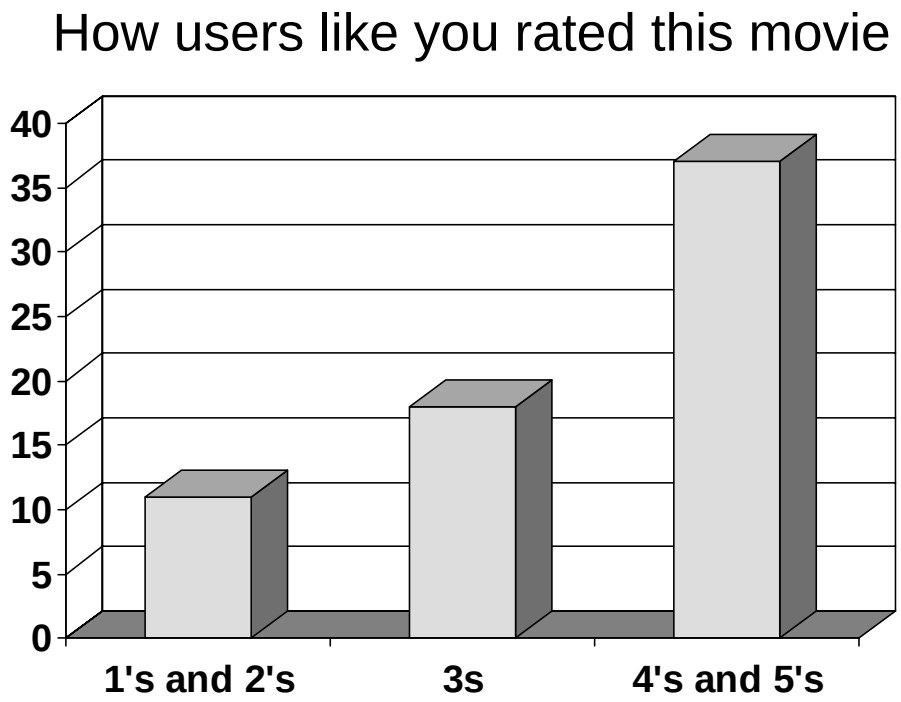
- Input:
  User Model, say in format (score, confidence)
  football (10,90%), cricket (1,92%),
  rugby (2,42%), tennis (6,72%)

- Output:
  System decides to emphasize football news,
  deemphasize cricket and rugby.

Example: AA Layer of Recommender

- Showing screen shots of how recommendations are explained

- Discussion on which way is preferable and how to improve



Focus Group – Summary

Advantages

- Can be done very early in the design process
- Can discuss events happening over long time span

Limitations

- Subjective opinions only:
  what people say they like might not be best for them
- Depends on good moderator
- Can only cover a few topics (or it will take too long)

Evaluation Methods

- Focus Group
- Cognitive Walkthrough
- Heuristic Evaluation
- Task-based Experiment
- Play with Layer
- User/Expert as Wizard
- Simulated Users

Cognitive Walkthrough – Traditional

- Uses usability experts, can be done early in design
- Focuses on learnability: ease of use for novice user

- Work through typical tasks,
  and decide for each task step
  whether a naïve novice user would have difficulty
  - » Will the user expect to do this
  - » Will the user see the control
  - » Will the user recognize the control is appropriate for the action step
  - » Will progress be apparent once the control has been used

Cognitive Walkthrough of an Adaptive System

- Best suitable for layers with GUI

- Apply Cognitive Walkthrough in the usual way
  - » **But** you typically need to look at multiple instances or action sequences

Example: Scrutability of UM

- Suppose there is a GUI for modifying UM
  - » Can apply Cognitive Walkthrough:
    will user be able to change the UM to a desired state

- Suppose there is no GUI for modifying UM directly
  - » Can apply Cognitive Walkthrough only if:
  - » There is a GUI to lower levels (e.g. GUI available for reading and rating news stories)
  - » The algorithm for the UM layer and the lower layers has been designed to the extent that a correct action sequence can be made (difficult when Machine Learning used)

Cognitive Walkthrough – Summary

Advantages

- Can be done early in the design process
- Task focus

Limitations

- Strongly GUI related
- Only considers learnability
- Certain adaptation aspects not covered
  - » Typically looking at the first time a user does a task, if done in the traditional way

Evaluation Methods

- Focus Group
- Cognitive Walkthrough
- **Heuristic Evaluation**
- Task-based Experiment
- Play with Layer
- User/Expert as Wizard
- Simulated Users

Heuristic Evaluation – Traditional

- Uses experts, can be done early in design

- Compare a system to a set of guidelines
- Often used in usability evaluation to compare a GUI to a set of guidelines
- Nielsen's 10 heuristics are frequently used
- Experts (3-5) work individually, results combined

Heuristic Evaluation of an Adaptive System

- Experts are given input like the layer would have, and shown what the layer does with it. They judge the layer's performance on a set of heuristics.
- Criteria discussed for layer can act as heuristics, but may need making more specific
- Could also use an adapted version of Nielsen's heuristics

Heuristic Evaluation of an Adaptive System

- Visibility of System Status (~Transparency)
  - » Does the user know what the system has interpreted and modelled?
    (ID: You spend 5 min reading this item; UM: You like cricket)
  - » Are adaptation decisions made visible to the user?
    (I will no longer show you football news)

- Consistency
  - » Is the adaptation not making the user experience too inconsistent?
  - » Can the user anticipate the system's adaptive behaviour?

Heuristic Evaluation of an Adaptive System

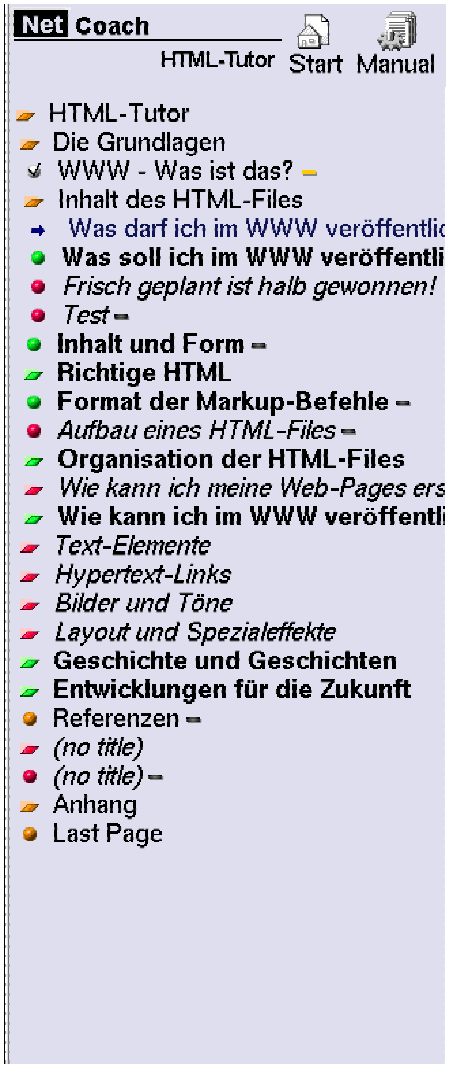
- User Control and Freedom (~Scrutability)
  - » Can the user undo a system interpretation?
    (I did not spend 5 min reading this news item, I went to the toilet)
  - » Can the user undo a modelling action?
    (I am not interested in cricket)
  - » Can the user undo an adaptation decision?
    (You will show me the football news!)
  - » Can the user decide e.g. when and how adaptations are applied?

Heuristic Evaluation of an Adaptive System

- Efficiency
  - » Normally, intended at expert users being efficient. Can look at this for some GUI related layers
  - » For many layers, it is not possible to judge this without looking at the algorithm (Could look at algorithm complexity)

- "Speaking the user's language"
  - » Are adaptations done in a way that fits with user's expectations from the real world?

Example: AA Layer of an ITS



- Are the symbols used for link annotation clear to the user (“speaking the users’ language”) ?
- Does the link annotation make the interface inconsistent?
- Is the system status visible?
- Can the user get rid of the link annotation?
- Can the user change the way link annotations are done?

Heuristic Evaluation – Summary

Advantages

- Can be done early in the design process
- Applicable more widely than GUI

Limitations

- Need to decide on appropriate heuristics
- Experts are not real users

Evaluation Methods

- Focus Group
- Cognitive Walkthrough
- Heuristic Evaluation
- **Task-based Experiment**
- Play with Layer
- User/Expert as Wizard
- Simulated Users

Task-based Experiment - Traditional

- Give user well-defined tasks to do
- Can measure time, errors, satisfaction, etc
- Observational Methods
  - » Thinking-aloud
    "Tell me what your are thinking"
  - » Co-discovery
    Two users do task together, and naturally talk
  - » Retrospective Testing
    Show video and ask what thinking at the time
  - » Coaching method
    Ask any questions to coach, learn what confuses

Task-based Experiment of an Adaptive System

- Particularly good for evaluating a set of layers (like the system as a whole)
- Can test how fast users find a book they like, how fast they learn, which adaptations they liked, which confused them

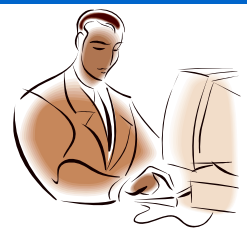
Problem:

- Adaptation takes time, can take too long for one session

Solutions:

- Longitudinal study: follow users over long time
- Focus on higher layers, with UM given (by or to user)

Example: Transparency of recommender UM



- System works at least up to the UM layer
- Transparency: do they understand how the modelling works

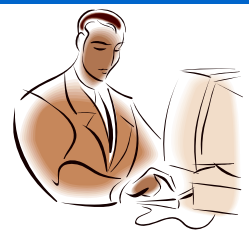
- For instance, can they get the system to believe they hate cricket and love football

- If direct interaction with UM is possible, can test scrutability

Example: DA and AA layers of ITS

- User is told to select a lesson to suit a learner, with characteristics of that learner (i.e. UM), and that the system knows these.
  (Allows focus on DA + AA layers.)
- Can measure e.g.
  - » Efficiency: how fast can the user decide?
  - » Effectiveness: is the user's decision the right one? (as judged by independent experts, having seen the lessons)
  - » Satisfaction: is the user pleased with their experience?
  - » Trust: does the user trust the system?
- Explain-your-decision question or Co-discovery

Task-based Experiment – Summary



Advantages

- Can be quite natural for users
- Can provide objective performance measures

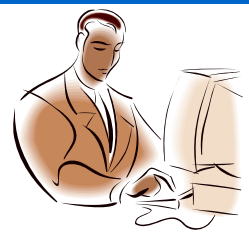
Limitations

- Requires the layer to have been implemented (or Wizard-of-Oz setup)
- Requires tasks that humans understand; easier for system as a whole, may be difficult for lower layers
- Input has to be easy to do:
  requires implementation of lower layers or special GUI
  OR: can tell them input, but then *indirect* experiment

Comment on indirect experiments

- In an *indirect* experiment, the user performs the task for somebody else, rather than for themselves
- We can control what kind of person they do the task for (give them UM)
- Helps to avoid time delay needed otherwise for adaptation…
- However, less natural for users and might make results less reliable

Comment on observational methods



- Normal lesson for experiments:
  - » Do not help the user! Let them struggle. (Unless coaching method used)
  - » Do not ask direct questions during the task, like what do you think of this label, as it may guide them
- However…. the user may not notice adaptation…
  - » it may be needed to interrupt them, and ask them about it explicitly
  - » e.g., if interested in scrutability, and they do not notice the scrutability tool, then might be good to lead them to it (but making a note to improve its visibility)
  - » or incorporate adaptivity-related activities in the tasks

Comment on observational methods (cont)

Normal limitations of observational methods:

- Thinking-aloud and Co-discovery interfere with users' cognitive processes, so can slow them down
- Thinking-aloud and Retrospective testing may lead to users justifying their errors, being insincere
- Users may not remember why they did things / what they thought afterwards (Retrospective testing)

In addition:

- Co-discovery may be less natural / suitable when a system is supposed to adapt to an individual user (unless user model provided)

Evaluation Methods

- Focus Group
- Cognitive Walkthrough
- Heuristic Evaluation
- Task-based Experiment
- Play with Layer
- User/Expert as Wizard
- Simulated Users

Play with Layer

Users or Experts test the layer by

- Freely inputting data as if coming from the layer below

Ways of evaluating layer:

- Judging whether the layer's behaviour is right on a set of criteria
- Questionnaire or interview to get user's opinions
- May also be able to get objective measures
  e.g., frequency of occurrences of certain events,
  like adaptation

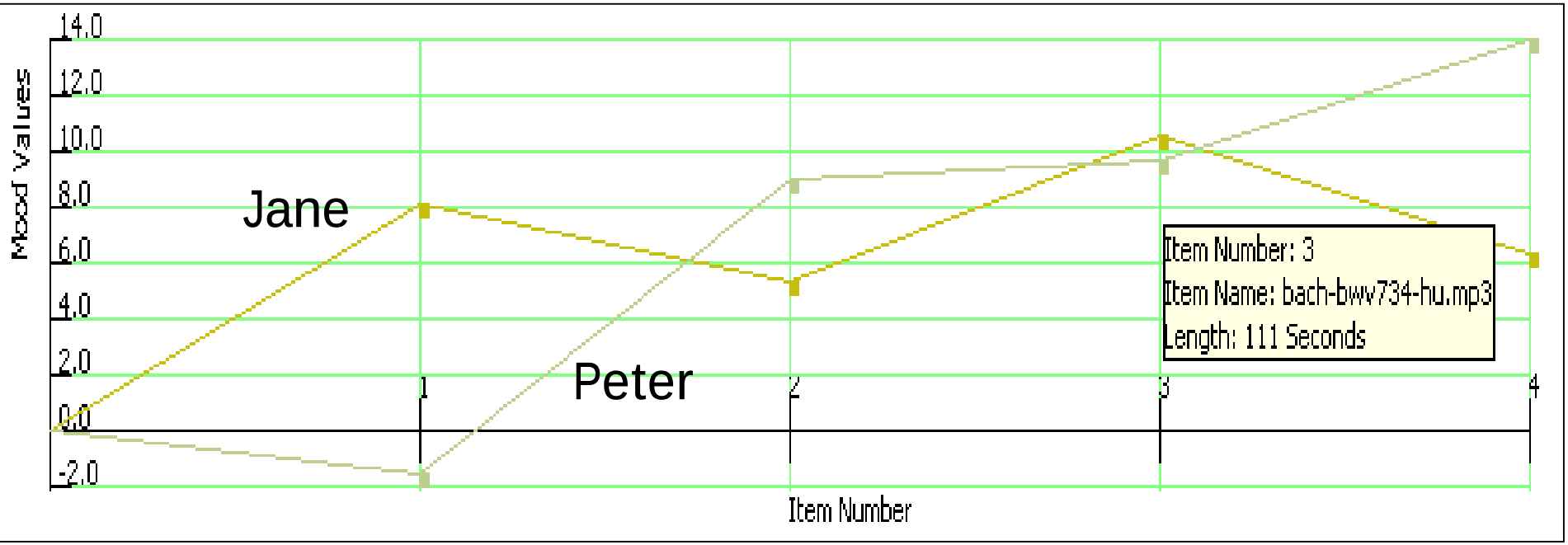
Example: CID and ID layers of Recommender

- Users can test out a CID layer which uses an eye tracker, by seeing how accurately and fast it picks up which movie they are looking at
- Requires an extra GUI element, showing the output of the CID layer

- Can also test ID layer, by telling users afterwards how interested the system thinks they are in each movie, based on what they were looking at
- May not require extra GUI, could replay interaction

Example: DA Layer of Group Recommender

- Simulator allows setting user profiles (ratings for music items), and simulating entry and exit of users from room



Example: DA/AA Layer of Recommender

- Users can set their own user model, via specially made GUI
- Explanations of recommendations are produced based on the UM
- Users rate the explanations on various criteria

Play with Layer

Advantages

- Can be done *before* the underlying layers have been implemented

Limitations

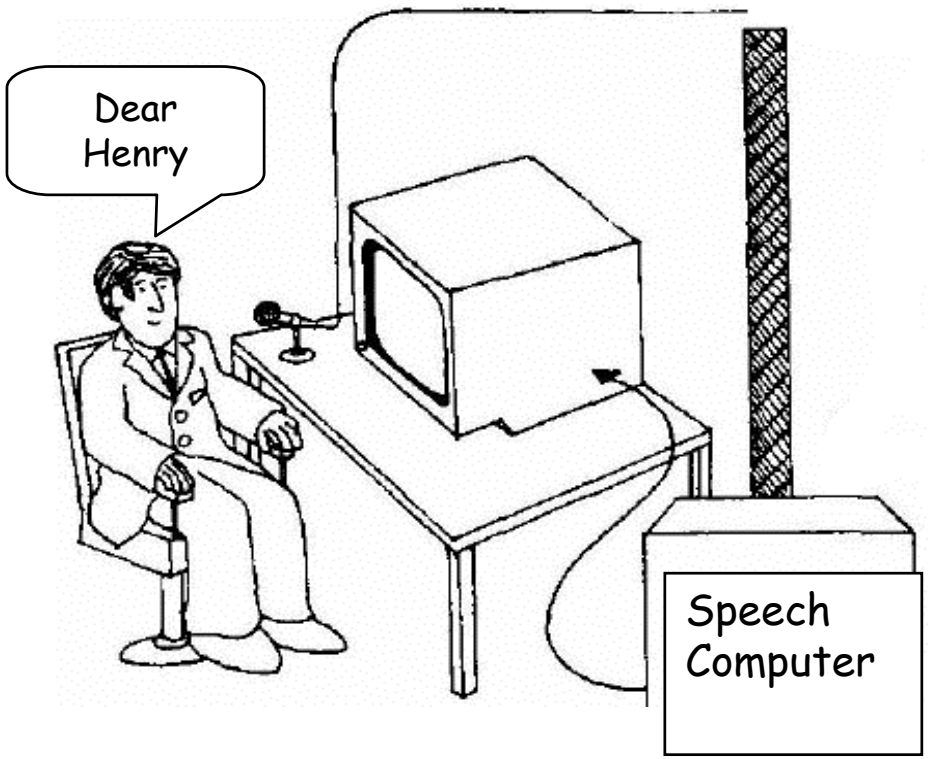
- Requires the layer itself to be implemented (though a Wizard-of-Oz could be used)
- Requires layer input to be understandable to and producible by the participant (difficult for a Bayesian UM)
- Likely to require a GUI for input (extra work)

Evaluation Methods

- Focus Group
- Cognitive Walkthrough
- Heuristic Evaluation
- Task-based Experiment
- Play with Layer
- User/Expert as Wizard
- Simulated Users

Wizard-of-Oz – Traditional

- Testing a non-existing system

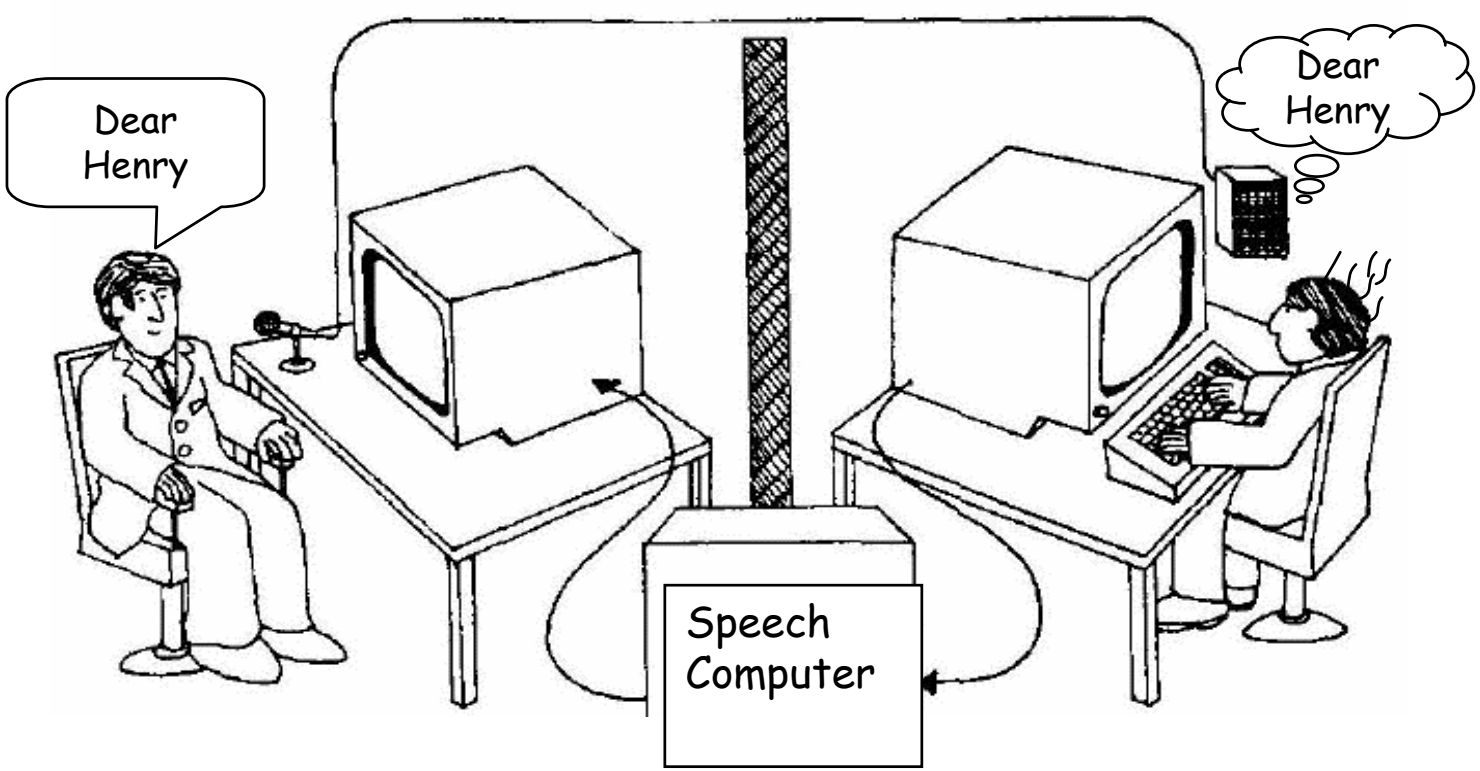


What the user sees

From Gould, Conti & Hovanvecz, Comm ACM 26(4) 1983.

Wizard-of-Oz – Traditional

- Testing a non-existing system



From Gould, Conti & Hovanvecz, Comm ACM 26(4) 1983.

User/Expert as Wizard (1)

- Participants are given input like the layer would have
- They perform the layer's task
  - » The same observational methods can be used as in a task-based experiment
- System performance is compared to their performance

Example: User Model Layer of a Persuasive System

Nuclear power has had 30 years of subsidies, billions of pounds poured into it and it still only produces 7% of the world's energy.

Position and strength verified in another experiment

I know Adam is neutral on Nuclear power. What will his position be after hearing this argument?

Why?

Compare what people do with what layer does

Example: AD Layer of a Group Recommender

	A	B	C	D	E	F
Peter	10	4	3	6	10	9
Jane	1	9	8	9	7	9
Mary	10	5	2	7	9	8

I know individual ratings of Peter, Mary, and Jane. What to recommend to the group? If time to watch 1-2-3-4-5-6-7 clips…

Why?

Compare what people do with what layer does

Example: DA/AA layer of a hierarchy generator

- Users given items, and asked to produce hierarchy
- Input: 37 items
  - *Discussion of the invention of antiseptics by Hungarian Ignaz Semmelweis in 1847.*
  - *Biography of Richard III who was king of England between 1483 and 1485.*
  - *Discussion of the play Henry VIII, published by Englishman William Shakespeare in 1623.*
  - *Biography of American Thomas Edison who invented the phonograph in 1877.*
  - *Discussion of the invention of the propeller by Englishman Francis Pettit Smith in 1835.*
  - *…*
- Co-discovery
- Compare what people do with what layer does

User/Expert as Wizard (2)

Alternative to final step:

- An expert review is conducted using the output of both participants and system *without* the experts knowing who produced which output

Some similarity with Turing test

Example: DA/AA Layer of a Hierarchy Generator

- Experts judged user and system generated hierarchies on a set of criteria
  (like understandability of titles, whether titles covered section content, etc)

  1. Biographies
     - 1.1 French royalty (3)
     - 1.2 English royalty (9)
     - 1.3 Painters and Inventors (4)
  2. Creations
     - 2.1 Important inventions (12)
     - 2.2 Paintings (6)
     - 2.3 Writings of Shakespeare (3)

- Why did you judge it this way?

Example: AD Layer of a Group Recommender

	A	B	C	D	E	F
You	10	4	3	6	10	9
Friend 1	1	9	8	9	7	9
Friend 2	10	5	2	7	9	8

You know the individual ratings of you and your two friends. I have decided to show you the following sequence. How satisfied would you be? And your friends?

Why?

User/Expert as Wizard – Summary

Advantages

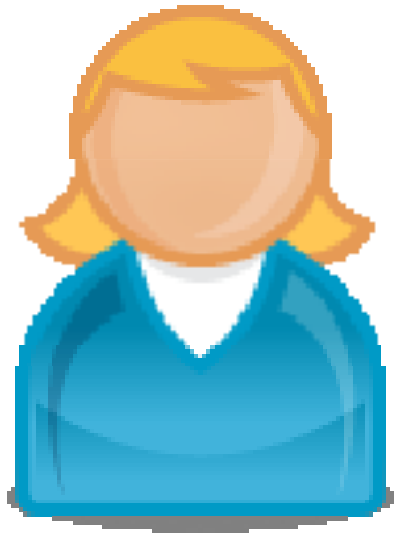
- Can be done *before* the underlying layers have been implemented
- Can even be done before the layer itself has been implemented
- May inspire design of the layer

Limitations

- Requires layer input to be understandable to the participant
- Requires task that humans are good at

Evaluation Methods

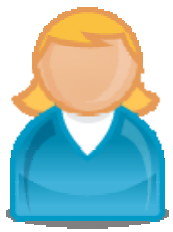
- Focus Group
- Cognitive Walkthrough
- Heuristic Evaluation
- Task-based Experiment
- Play with Layer
- User/Expert as Wizard
- Simulated Users



Simulated Users – Traditional

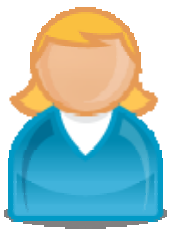
- Using real users in experiment costs time and money
- Difficult to control real users
  (e.g. if I want to test out many different types of users, how to make sure I get all these types)

- Use simulations of users instead of real users
  - » Theoretical approaches like GOMS (e.g. theory on how long it takes normal user to move mouse, press button, etc)
  - » Implementations, e.g. neural networks or probabilistic models



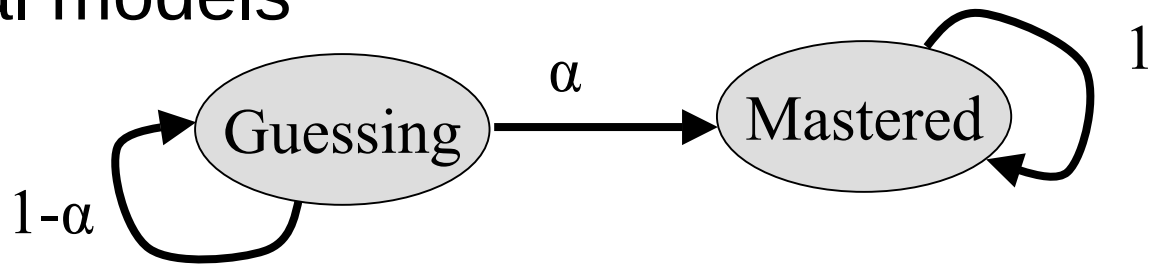
Simulated Users for Adaptive System

- Adaptive system requires many different types of user (point of adaptation!)
- Additionally, difficult to get input for layer, e.g. want to test DA layer, but how to get exact UM from users
- Test the layer using simulated users



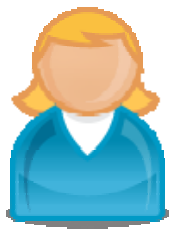
Example: Simulated Users for ITS

- ITS for teaching paired associates (Japanese translations of Dutch words)
- Considered several models
  e.g., All-or-None (Bower, 1961)



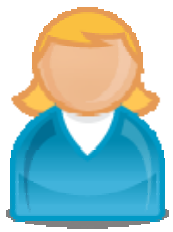
P (Correct response | in Mastered)=1
P (Correct response | in Guessing)=g

- Models predicted how well variants of DA layer do
  (how many correct responses simulated users get on average for three strategies that select items to learn)



Example: DA layer of Group Recommender

- Simulated users in terms of affective state
- Affective state models contained two parameters (between 0 and 1), users likely to vary, and not clear what good values in general
- Ran simulations with all kinds of values
- Looked at how (un)happy simulated users would be with output from different variants of DA layer
- Learned what variants did not work, independent of value parameters



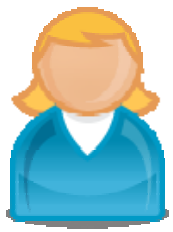
Simulated Users – Summary

Advantage

- Can test many things quickly

Limitations

- The models used for the simulated users are likely to be based on the same assumptions that underlie the adaptive system's design. What if those assumptions are wrong?
- Modelling static user behaviour differs from modelling *adaptive* user behaviour



Evaluation Methods Overview

Evaluation Method	Where in Design Process	For which layers in particular
Focus Group 	Requirements, Design	UM, DA, AA
Cognitive Walkthrough 	Design (of GUI)	DA+AA, Complete System
Heuristic Evaluation 	Design	Any Layer
Task-based experiment 	Implementation	DA, AA, Complete System
Play with layer 	Implementation	Any Layer
User / Expert as Wizard 	Requirements, Design (of Alg.)	UM, DA, AA
Simulated users 	Design	DA, AA

Evaluation Methods Summary

- Task-based experiments are just ONE method for evaluating adaptive systems, many others exist
- Good to evaluate early on in the design, not just at the end
- Formative aspect of evaluation is important: not just how good it is, but what causes problems
- Skill required to evaluate a layer separately, shown you examples of how to do this
- Best method depends on the type of system and when the evaluation is happening
- *Traditional methods may need to be adapted to suit the requirements of adaptive systems*

“Hands-on” Session

2nd Part

Creating a concrete evaluation plan – Selecting evaluation methods and data collection instruments

“Hands-on” session overview – 2nd Part

- Goals for this part
  - » Select evaluation methods and data collection instruments
  - » Relate these to the output of the first session (e.g., what methods for what criteria)
  - » Understand how to lay out an evaluation plan based on the above
- Organisation of second part
  - » Re-establish the first session’s groups
  - » Group discussion / work
  - » Sampling / presentation of results

“Hands-on” session overview – 2nd Part

- Expected output
  - » A list of evaluation methods that you would use for the system at hand
    - Including what data collection techniques you would employ
  - » A “cross-reference” between the layer- and domain- specific evaluation topics and criteria from the first session, and the evaluation methods / instruments
  - » *Optionally* an outline of a time- / sequence- plan for the evaluation

- Time available: 30 minutes

Questions to focus on

- First step
  - What evaluation methods would be a “best fit”?
  - Same for data collection methods
- Second step
  - How can everything be put together to create a coherent evaluation plan?
  - How can an adaptivity-oriented evaluation plan be reconciled with more “traditional” HCI oriented ones?

Pitfalls

Common mistakes to avoid

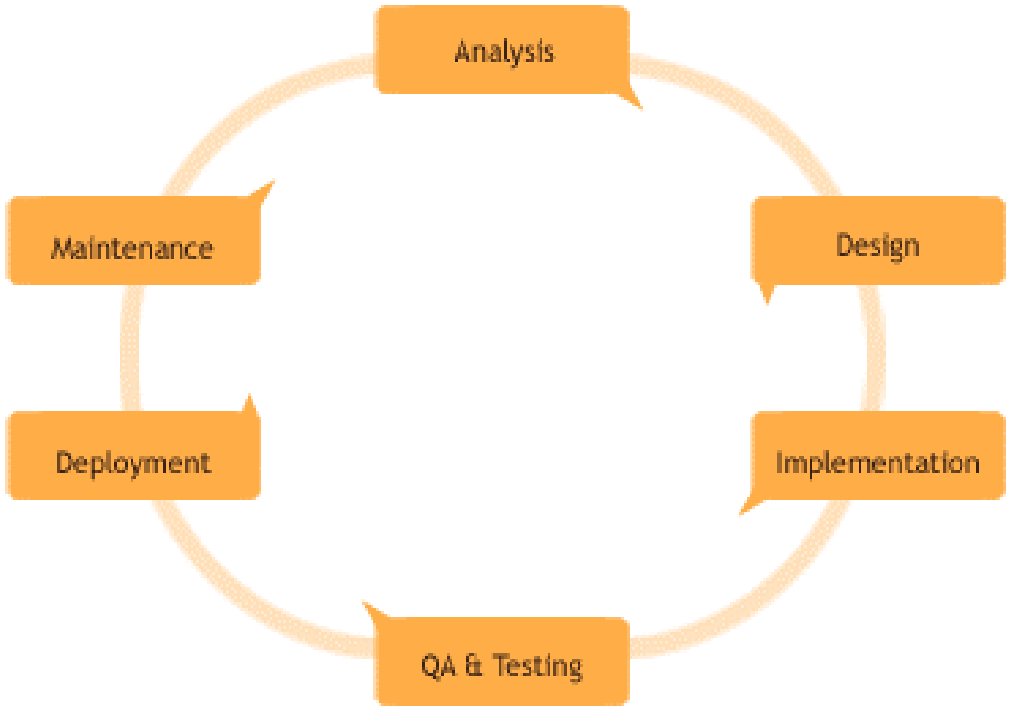
Which pitfalls should I try to avoid?

1. Big evaluation study planned for the end of a project
2. Not enough resources left
3. Wrong control condition selected
4. Too much variance in data
5. Confusion which criterion to choose
6. Users are unable to tell about adaptivity effects
7. Evaluation results are reported incomplete or anecdotally

Graphic © Video Game Critic, 2007

Pitfall 1: Big evaluation study at project end

- Big evaluation study planned for project end
- Summative evaluation cannot recover failures in earlier stages
- Recommendations
  - » Conduct several formative studies (cf. methods section)
  - » Distributed across the development cycle



Graphic © GNSE Group, 2007

Pitfall 2: Not enough resources left

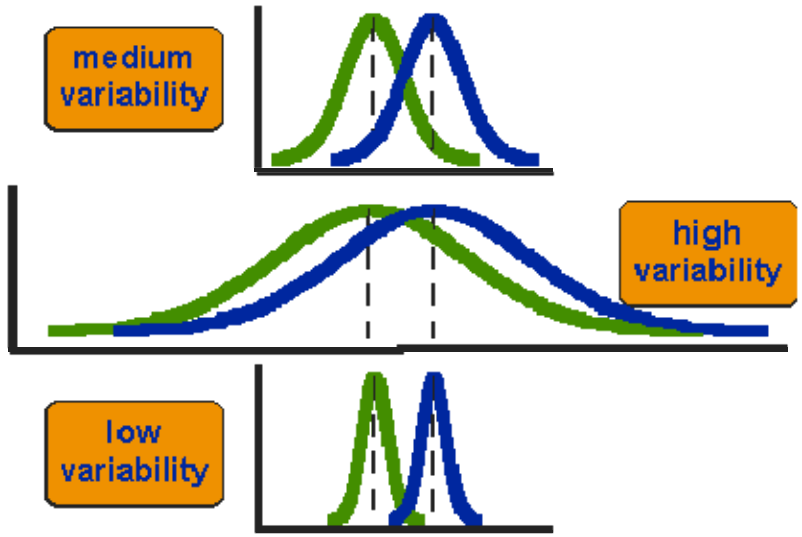
- Empirical studies require personnel, organizational, and financial resources
- Recommendations
  - » Spread studies across the development cycle
  - » Expert evaluation
  - » Evaluate inference mechanism with simulated users and empirical data
  - » Use cognitive models

Pitfall 3: Wrong control condition selected

- Switching off the adaptivity might result in an incomplete or even useless system
- Recommendation
  - » Compare various adaptation decision conditions that are based on the same user characteristics

Pitfall 4: Too much variance in data

- High variance corrupts statistical analysis
- Recommendations
  - » Try to find a sample that is
    - **heterogeneous** in terms of the modeled user characteristics,
    - but **homogeneous** in terms of other factors
  - » Use repeated measurement
  - » Control variables that might have an impact on the results
  - » Separate groups of users



Graphic © Research KB, 2007

Pitfall 5: Confusion which criterion to choose

- There is no single evaluation criterion for adaptivity
- Recommendations
  - » Define goals of adaptivity precisely
  - » Derive criteria from these goals

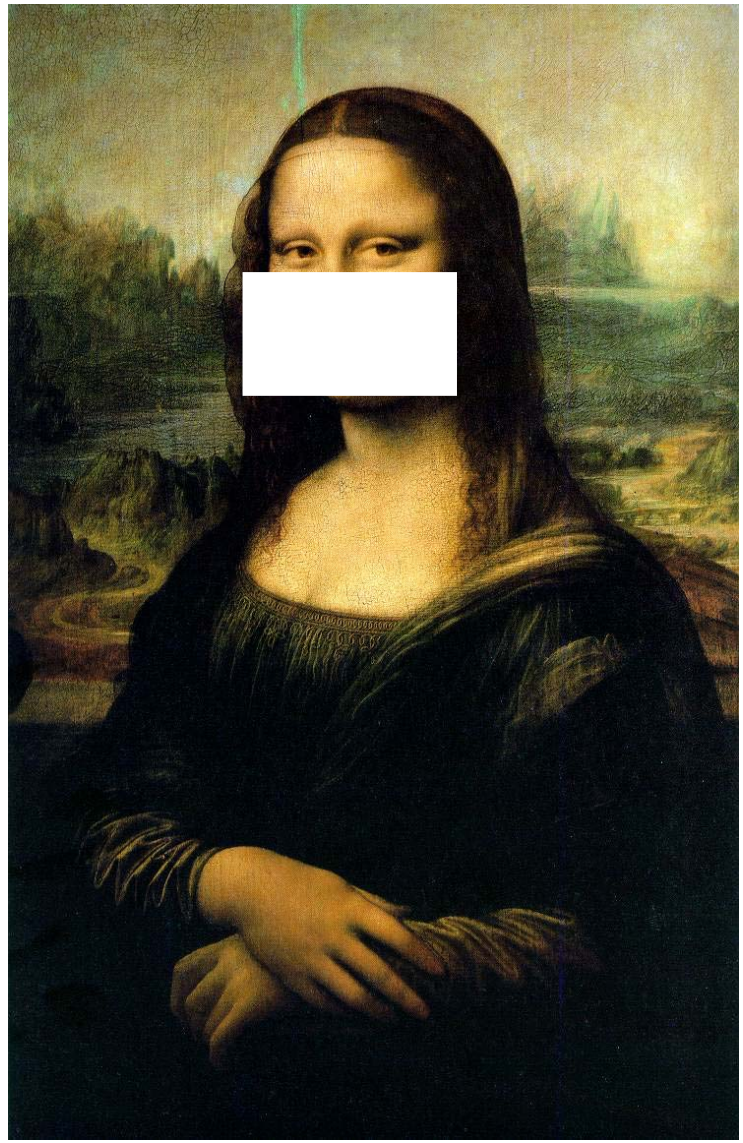
Photo © Sport Thieme, 2007

Pitfall 6: Reporting Adaptivity Effects

- Users might be unable to tell about adaptivity
- Users might not have noticed adaptivity at all
- Recommendations
  - » Use user feedback in combination with objective measures
  - » E.g., log-files, behaviour observation

Pitfall 7: Results Reported Incomplete

- Results are often reported incomplete or anecdotally
- Incomplete report of results corrupts interpretation of study
- Recommendations
  - » Guidelines on reporting statistical data
  - » Include important information for adaptivity (e.g., empirically identified user characteristics, effect size)



Recommendations

- Plan carefully and in advance
  - » Sample
  - » Control condition
  - » Criteria
- Slice (or dice) system: Layered Evaluation
- Publish your results

Where to next?

Reading List

- [User as Wizard] Nguyen, H., Masthoff, J. & Edwards, P. (2007). Modelling a receiver's position to persuasive arguments. Proc of the Persuasive Conference (Stanford, USA).

- [Usability methods] Gena, C. & Weibelzahl, S. (2007). Usability Engineering for the Adaptive Web. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): The Adaptive Web: Methods and Strategies of Web Personalization, Lecture Notes in Computer Science, Vol. 4321 (© Springer). Berlin: Springer.

- [User as Wizard, Simulated Users] Masthoff, J. & Gatt, A. (2006). In pursuit of satisfaction and the prevention of embarrassment: Affective state in Group Recommender Systems. *User Modeling and User Adapted Interaction, 16*, pp281-319.

- [User as Wizard] Masthoff, J. (2006). The user as wizard: A method for early involvement in the design and evaluation of adaptive systems. Proc of the Fifth Workshop on User-Centred Design and Evaluation of Adaptive Systems, held at AH'06 (Dublin, Ireland).

Reading List (cont)

- [Pitfalls] Weibelzahl, S. (2005). Problems and pitfalls in the evaluation of adaptive systems. In S. Chen & G. Magoulas (Eds.). Adaptable and Adaptive Hypermedia Systems (pp. 285-299). Hershey, PA: IRM Press.

- [Layered Evaluation] Paramythis, A. & Weibelzahl, S. (2005). A Decomposition Model for the Layered Evaluation of Interactive Adaptive Systems. In Ardissono, L., Brna, P., & Mitrovic, A. (Eds.), Proceedings of the 10th International Conference on User Modeling (UM2005), Edinburgh, Scotland, UK, July 24-29 (pp. 438-442) (Lecture Notes in Computer Science LNAI 3538, Springer Verlag). Berlin: Springer.

- [Task-based Experiment, Simulated Users] Masthoff, J. (2002). The evaluation of adaptive systems. In N. V. Patel (Ed.), *Adaptive evolutionary information systems*. Idea Group publishing. pp329-347

- [Task-based Experiment] Weibelzahl, S., & Weber, G. (2002). Adapting to prior knowledge of learners. In de Bra, P., Brusilovsky, P., & Conejo, R. (Eds.), Proceedings of the second international conference on Adaptive Hypermedia and Adaptive Web Based Systems, Malaga, Spain, AH2002 (pp. 448-451). Berlin: Springer.

Reading List (cont)

- [Empirical Evaluation] Weibelzahl, S., Lippitsch, S., & Weber, G. (2002). Advantages, opportunities, and limits of empirical evaluations: Evaluating adaptive systems. Künstliche Intelligenz, 3/02, 17-20.

- [Layered evaluation] Weibelzahl, S. (2001). Evaluation of adaptive systems. In M. Bauer, P. Gmytrasiewicz & J. Vassileva (Eds.), User Modeling 2001: Proceedings of the Eighth International Conference, UM2001. (pp. 292-294) (Lecture Notes in Computer Science LNAI 2109; © Springer-Verlag). Berlin: Springer.

- [Layered evaluation] Paramythis, A., Totter, A., & Stephanidis, C. (2001). A modular approach to the evaluation of Adaptive User Interfaces. In S. Weibelzahl, D. Chin & G. Weber (Eds.), Proceedings of the Workshop on Empirical Evaluations of Adaptive Systems, held in the context of the 8th International Conference on User Modeling (UM'2001), 13-17 July, Sonthofen, Germany (pp.9-24). Freiburg: Pedagogical University of Freiburg.

- [Task-based Experiment] Chin, D.N. (2001). Empirical Evaluation of User Models and User-Adapted Systems. User Modeling and User-Adapted Interaction 11: 181-194, 2001.

- [Pitfalls] Höök, K., Karlgren, J., Waern, A., Dahlback, N., Jansson, C., Karlgren, K., and Lemaire, B. (1996). A glass box approach to adaptive hypermedia. User Modeling and User-Adapted Interaction, 6:157-184, 1996.

Additional resources

- Tutorial's site
  http://www.easy-hub.org/hub/tutorials/um2007/
  - » Expanded "reading list" for this tutorial
- Evaluation of Adaptive Systems Hub
  http://www.easy-hub.org/
  - » Previous workshops
  - » Guidelines
  - » Literature references