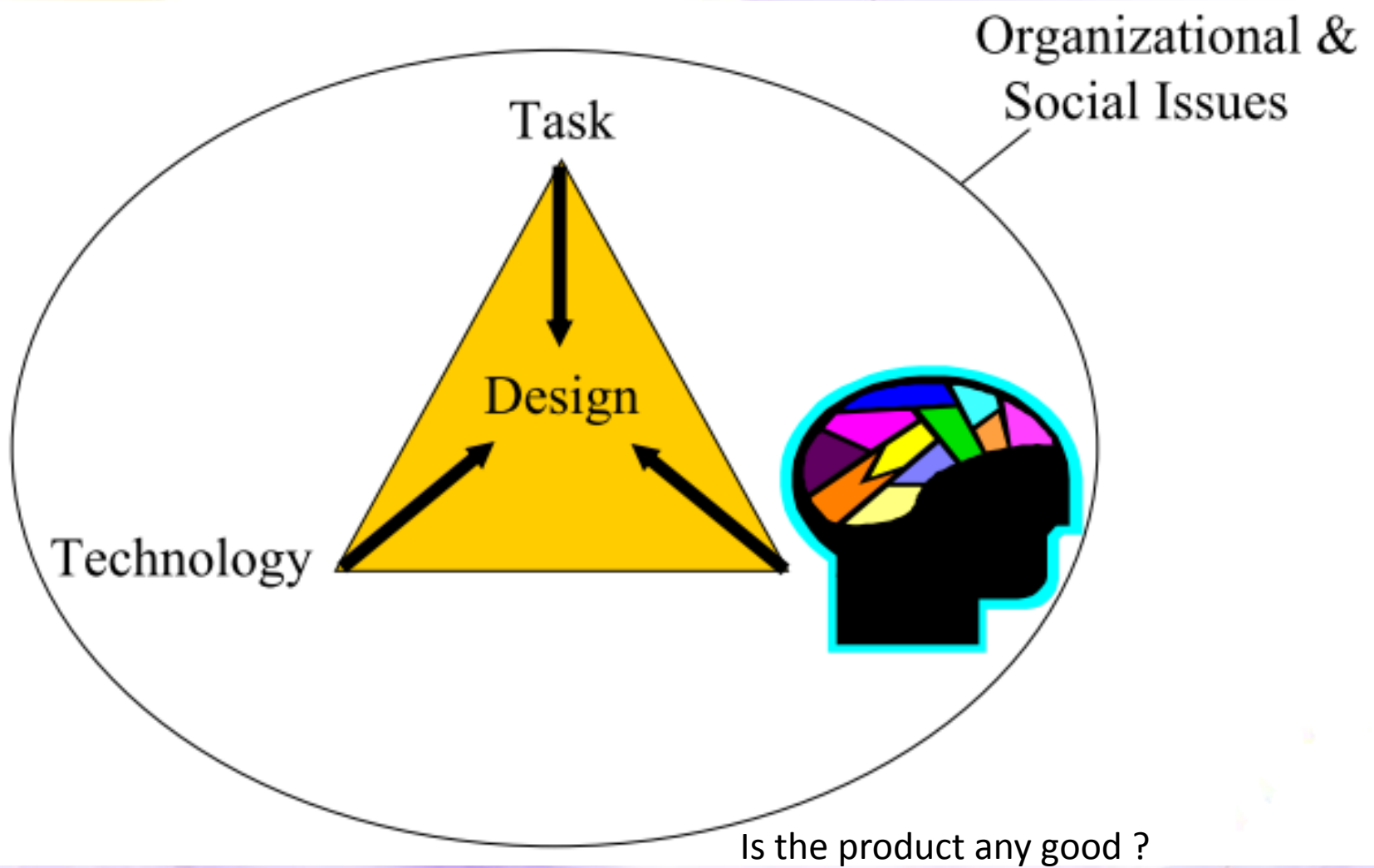
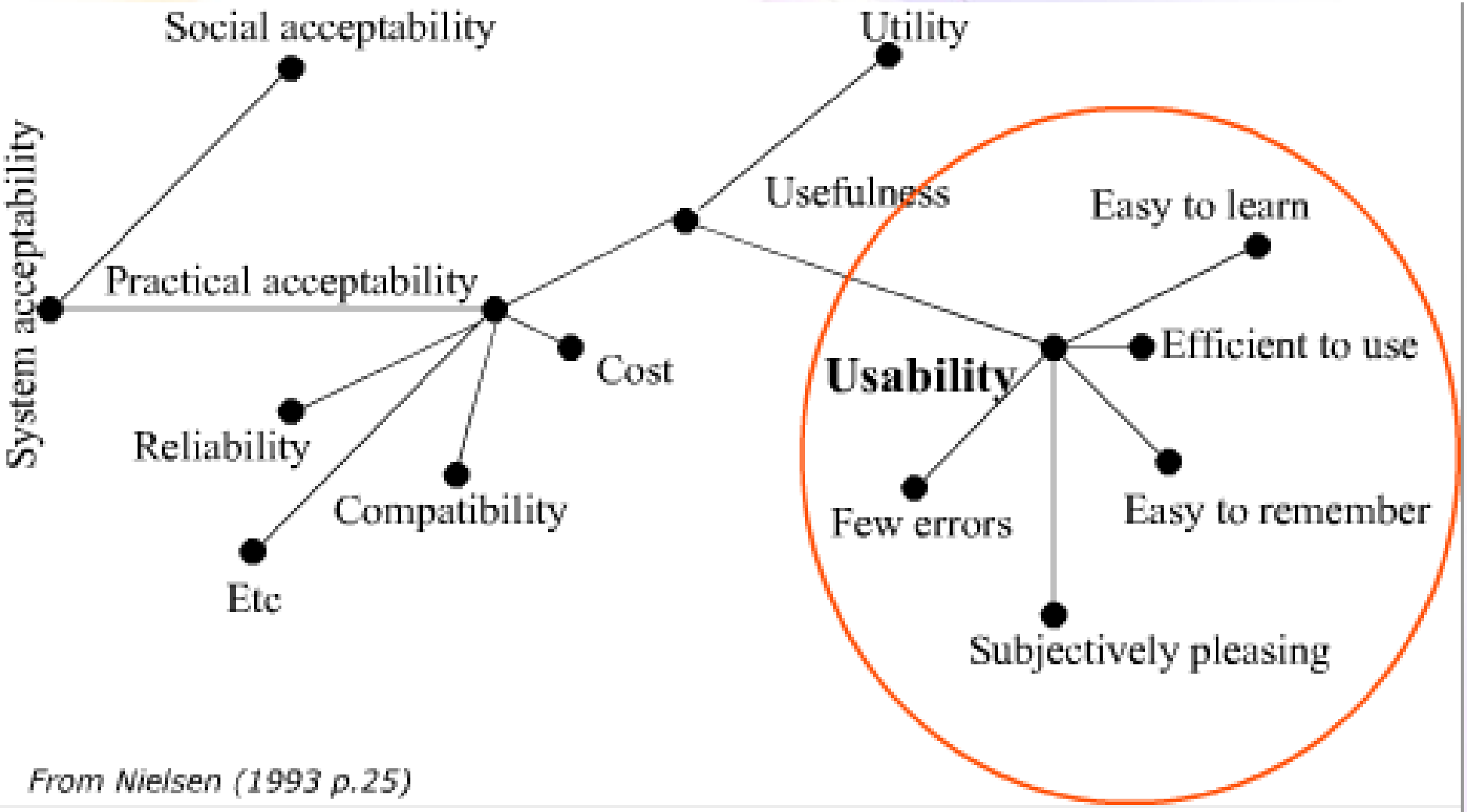
Testing & Evaluation



Is the product any good ?

Conduct tests with the real users

System acceptability revisited



From Nielsen (1993 p.25)

Usability evaluation OR Functionality testing

- Testing: test functionality of system identify & fix bugs, faulty logic etc

  → Clearly important
- Evaluation: test usability of system can users achieve their goals in terms of:
  - Effectiveness
  - Efficiency
  - Productivity
  - Safety
  - User satisfaction

Test plan

- Purpose of test
- Problem statements/test objectives
- Participant profile (inclusion/exclusion criteria)
- Method/technique to be used
- List of tasks to be used
- Test environment (field vs. lab) and material (HW/SW, resources → recorder and batteries, report forms, questionnaires)
- Experimenter's role (monitor, coach etc.)
- Evaluation measures to be taken (qualitative vs. quantitative, subjective vs. objective)
- Contents of report to be produced and how the report is going to be presented → focus group, informal meeting, big bosses are going to be there

Evaluation paradigms

Quick and dirty: informal discussions with users at any time perhaps using prototypes.

Get feedback from users or consultants to confirm that their ideas are inline with users' needs and are liked.

Field studies: going to the users' sites and using surveys or observing users using the UI

What users do naturally and how technology impacts them

Usability testing: observe the user and record typical users' performance on typical tasks in controlled settings

Explain why the users did what they did (calculate performance times, identify errors)

Elicit users' opinions (interviews, questionnaires)

Predictive: users need not be present (generally at the developers site)

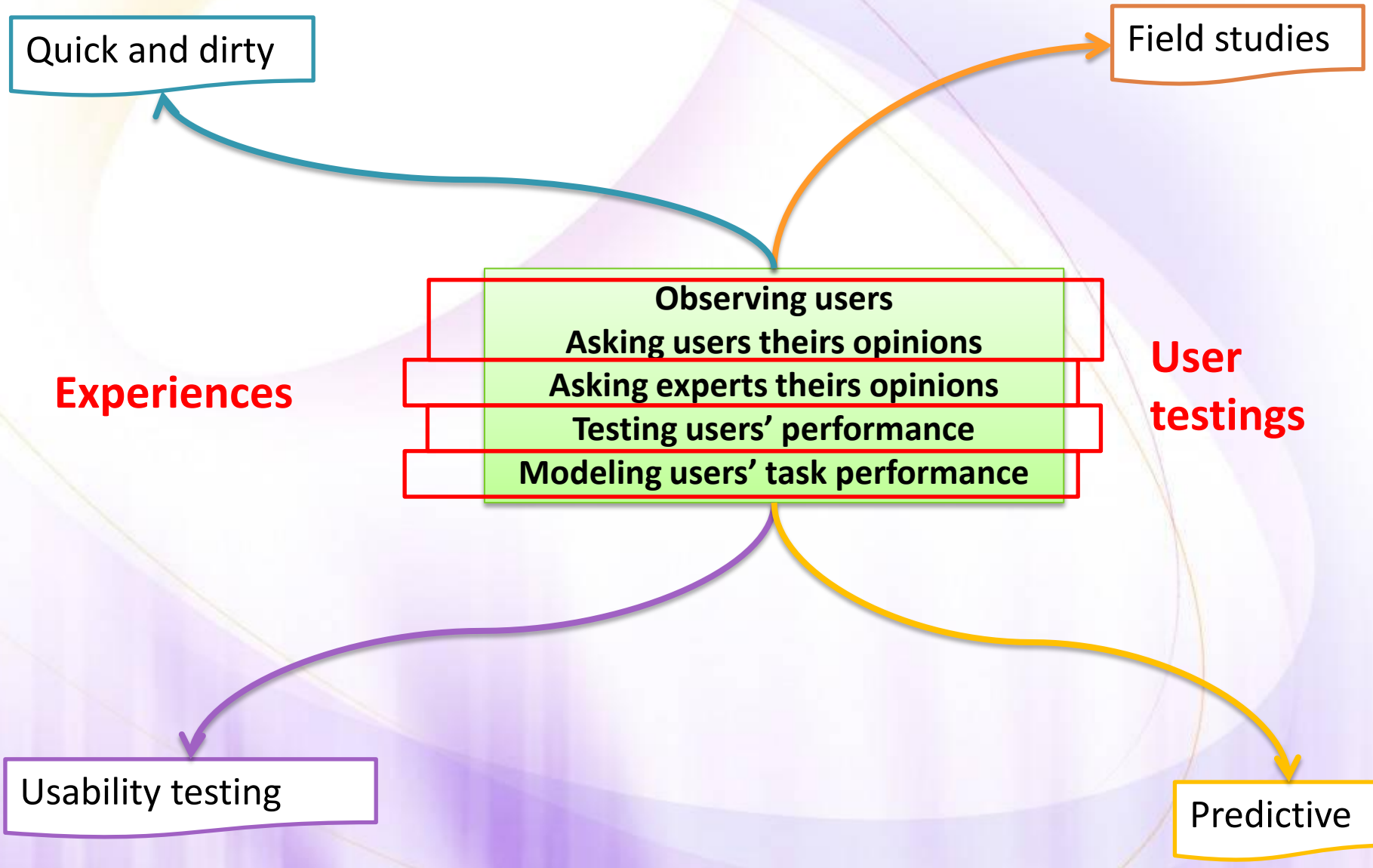
Experts apply their knowledge of typical users (using heuristic evaluations) to predict usability problems.

Another approach involves theoretically based models.

Differences between evaluation paradigms

Evaluation paradigm:	Quick and dirty	Usability testing	Field studies	Predictive
Role of the user in the evaluation	Natural behavior	To perform tasks	Natural behavior	None
Who controls the evaluation	Evaluators has minimum control	Evaluator has strong control	Relationship between Evaluators and customers	Expert Evaluators
Location of the evaluation	Natural environment or lab	Lab	Natural environment	Lab or on the premises
When the evaluation **is used**	Any time	with prototype or product	Early	with prototype
Type of data collected from the evaluation	qualitative; informal discussions	Quantitative; statistical	Qualitative, sketches	List of problems
How the data is **fed back** into the design	Sketches and quotes	Report on performances	Descriptions at workshop, reports and sketches	Report
Philosophy or theory of the evaluation	User-centered	Scientific/experimental	Ethnographic	Practical heuristics, Theory

Evaluation paradigms



Mappings

	Quick and dirty	Usability testing	Field studies	Predictive
Observing Users	Seeing how users behave in their environment	Video and interaction logs. Analyzed for errors, performance, route in UI, etc	Ethnography is central to Field Studies	NA
Asking Users	Discuss w/ potential users, individually or in focus groups	Pre and Post testing surveys. Structured Interviews	Interviews or discussions	NA
Asking Experts	Provide usability critiques on prototypes	NA	NA	Heuristic Evaluation
User testing	NA	Testing typical users on typical tasks. Central to UT	NA	
Modeling	NA	In academia compare w/ theory.	In academia compare w/ theory	GOMS etc

Covered topics

- Experience
- User Testing

Classification

- Formative evaluation: before implementation
  - helps to `form' ideas
  - part of iterative design/RAD
  - very difficult to manage
- Summative evaluation: at the end of development
  - `9 out of 10' cats preferred it;
  - difficult to make sense of results
- Post-summative evaluation: after implementation
  - monitor acceptance

Evaluation techniques

- Formative evaluation
  - cognitive walkthrough
  - pencil and paper mock-ups
  - cooperative evaluation
  - scenario based evaluation

→Low cost, do it often

→Find usability problems EARLY

- Summative evaluation:
  - lab-based techniques
  - logging and tracking
  - ethno-methodology

→ Mostly high cost, focused

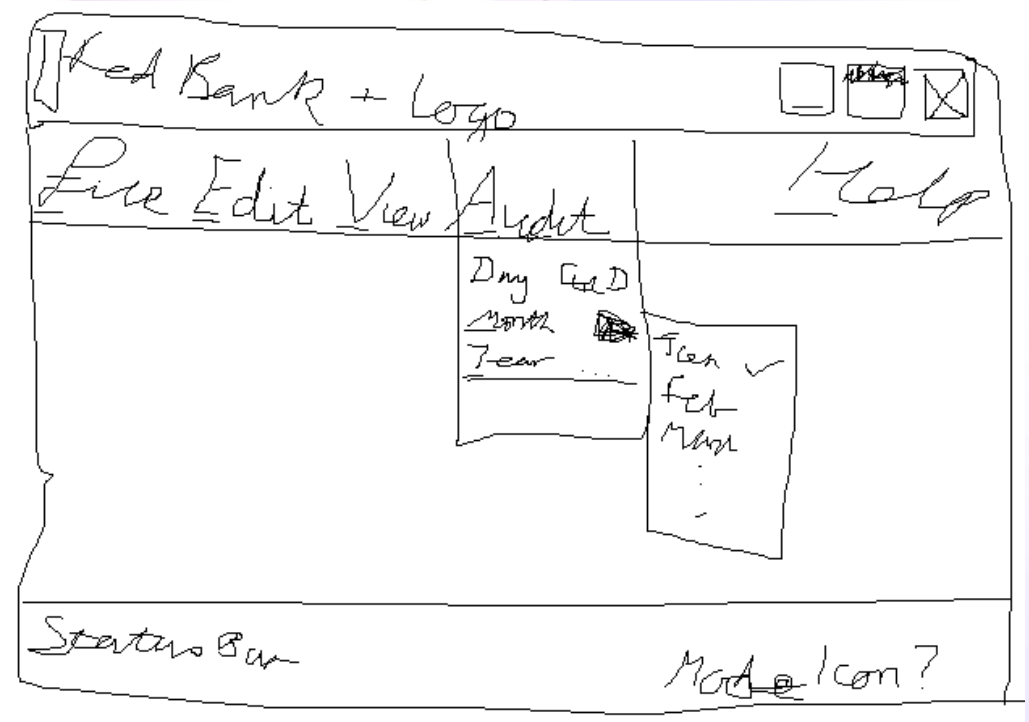
→ Demonstrate usability at the end.

- Post-summative Evaluation: If you don't spot problems:
  - huge and recurring training costs;
  - large amounts of documentation;
  - high staff turnover;
  - large numbers of support staff;
  - eventual need for re-design.

→ Importance of interface design.

Formative Evaluation : Pencil and Paper Mock-Ups

- Start with OHP foils:
  - overlay them to show menus etc.
  - BUT user bias against low fidelity?
- Move to higher fidelity:
  - BUT be ready to throw it away
  - whole point is to test assumption
- Get users involved early:
  - don't trust designer's opinions...



Formative Evaluation : Cooperative

- Get users to look at prototype:
  - DONT ask them what they think.
  - Do get them to perform a task
- Ask them to `think aloud':
  - If they say they're confused, note this down for later re-design.
- Where do you do this?
  - final environment?
  - or your design office?

Formative Evaluation

Walkthrough

- First domain expert:
  - identifies critical tasks
  - sets context of task
  - assumptions about users
- Walk through' task:
  - predict pitfalls for the user
  - use these in design/tests.
- Cognitive walkthroughs:
  - classify mental activities
  - `premature loss of goal'
  - `supergoal kill-off’

Scenario-based

- Users describe their `dream':
  - ideal trace of interaction;
  - like a transcript from a play.
- Designer takes this away:
  - forms a series of test cases;
  - the user must be able to…
- Good for contracts:
  - knows what user/system must do...

Summative Evaluation: Lab-Based Techniques

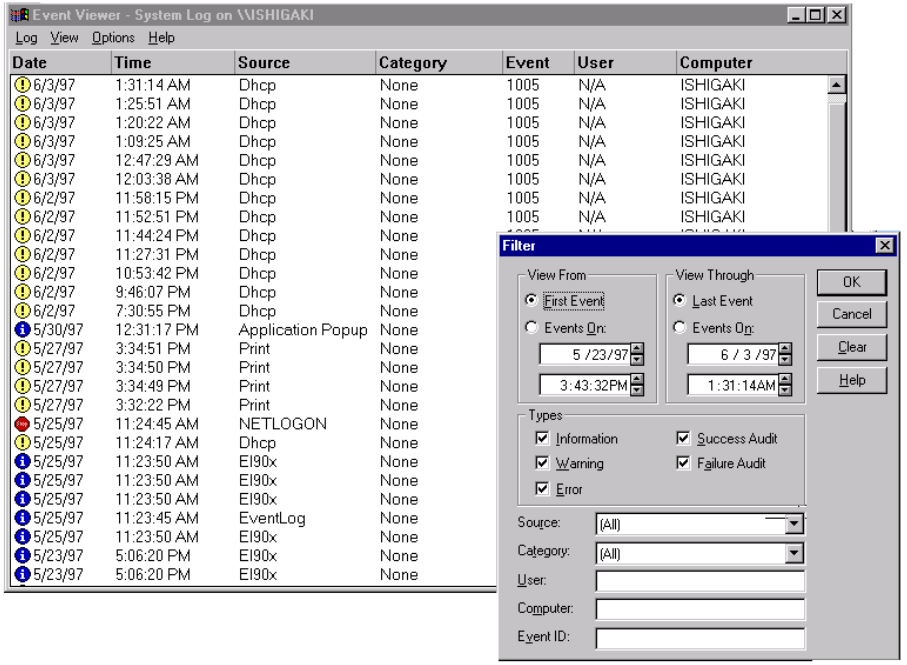
- Test specific hypothesis;
- Fewer errors on a set task;
- Faster entry of same data.
- Problems:
  - how representative are the tasks?
  - how representative are the users?
  - how representative is the context?

- Usability labs:
  - user out of normal environment.

- Aids comparison:
  - users get same interruptions.

Summative Evaluation: Logging and Tracking

- Monitor a few users with system in working environment
- Can record speed, errors etc.
- Problems:
  - what do you do with mass of data?
  - tell users about monitoring?
  - any resources left to fix problems?



- Is this legal or ethical?

Comparisons

- Empirical tests:
  - evaluate the evaluation methods
- As you'd expect:
  - different strengths/weaknesses;
  - lab-based less cost effective;
  - walkthroughs less coverage.
- Doesn't matter which you do but
  - do something....

Covered topics

- Evaluation
  - Experience
  - User Testing

Why do User Testing?

- Can’t tell how good UI is until?
  - People use it!
- Other methods are based on evaluators who
  - may know too much
  - may not know enough (about tasks, etc.)
- Hard to predict what real users will do

What will you mesure ?

- Time on Task
  - How long does it take people to complete basic tasks?
  - For example, find something to buy, create a new account, and order the item.
- Accuracy
  - How many mistakes did people make?
  - And were they fatal or recoverable with the right information?
- Recall
  - How much does the person remember afterwards?
- Emotional Response
  - How does the person feel about the tasks completed?
    - Confident?
    - Stressed?
    - Would the user recommend this system to a friend?

Usability Evaluation Testing

- Involving users directly in usability laboratory tests
- Thinking Aloud Protocol
  - Users are asked to verbalize their thoughts, feelings and opinions when using the system.
  - Two variations: critical response & periodic report.
  - Useful to understand user's mental model, interaction with the system and terminology.
- Co-discovery Learning
  - Users work in pair to achieve a common goal with the tested system.
  - Users are encouraged to "think-aloud".
- Coaching Method
  - Participants asked an expert coach.
  - Participant-coach & participant-computer interactions are observed.

Usability Evaluation Testing

- Shadowing method
  - An expert user sits next to a user and explains the user's behaviour to the usability tester.
  - Usually when it is not appropriate for users to give any response during the test session.
- Teaching method
  - User interacts with the system to gain some expertise.
  - A novice user is brought in, and the "expert" user is asked to explain to the novice user how to perform the task.
- Performance Measurement
  - Obtaining quantitative data when participants perform a task.
  - Minimize participant-tester interaction during the test as it might affect quantitative data.
  - Careful design of number of participants to ensure enough statistical power.

Usability Evaluation Testing

- Question-asking Protocol
  - Tester asked questions directly to participants.
  - Participants asked in terms of their past experience of in relation to the tested system.
- Remote Testing
  - Tester are separated in time/space from users.
  - Data are obtained from logs/records/networks.
- Retrospective Testing
  - Tester & participants reviewed recorded session together, tester asked what was going on.
  - Should be used together with other methods.

Practices : Usability Testing Report

- **Cover page**: name the group and the usability expert
- **Introduction**: Description of UI, test goals and brief description of tests (1 page)
- **Test Plans**: the original you created for testing (approximately 1 page per scenario)
- **Results**: These should be plots and charts with explanations (what it takes, ~ 2 pages)
- **Conclusions**: Usability problems and suggestions for improving the UI (1 page)