



# Effective Evaluation

Dr. Fotis Spyridonis, Dr. Mary Kiernan



# Structure of this lecture

- Understanding of key concepts and terms used in evaluation
- Introduction to different types of evaluation methods and how they are used
- Understanding of usability and its importance
- Overview of the typical usability testing process and tools

# Learning Outcomes

- Evaluation techniques are explored and this links to the learning outcome "Understand the issues involved in developing and **evaluating** interfaces to interactive applications".

# Understanding evaluation

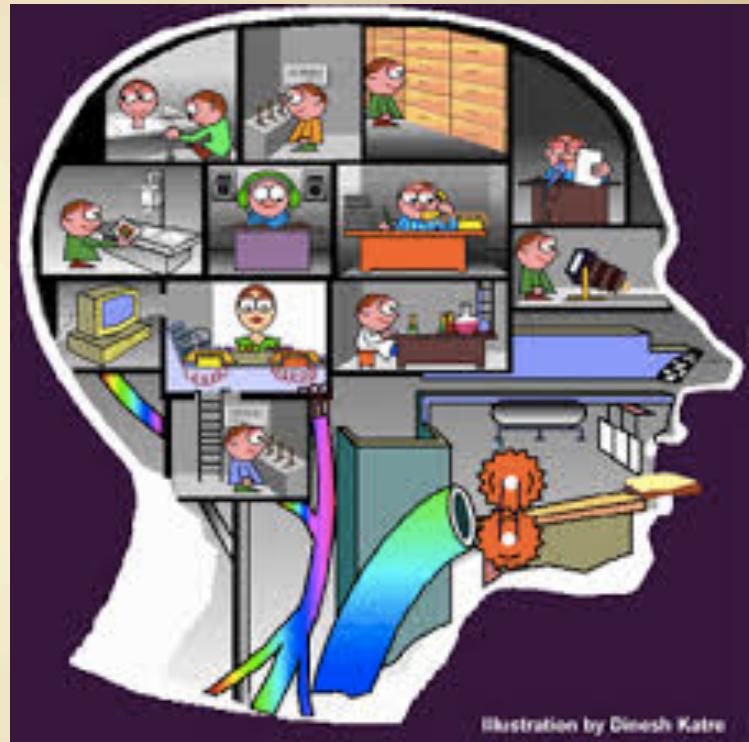
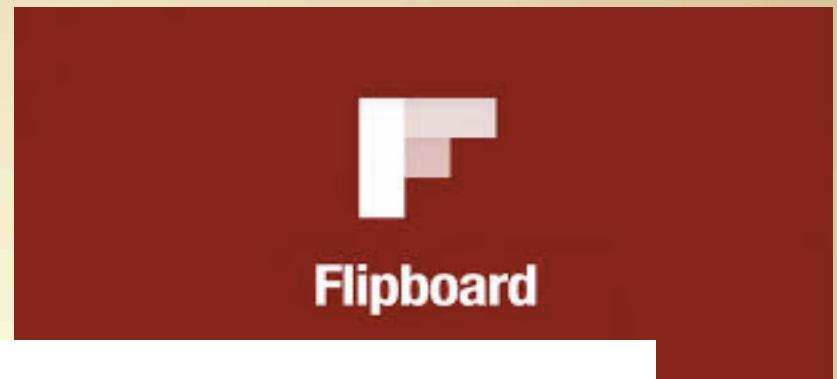


Illustration by Dinesh Katro

# The age of user experience...

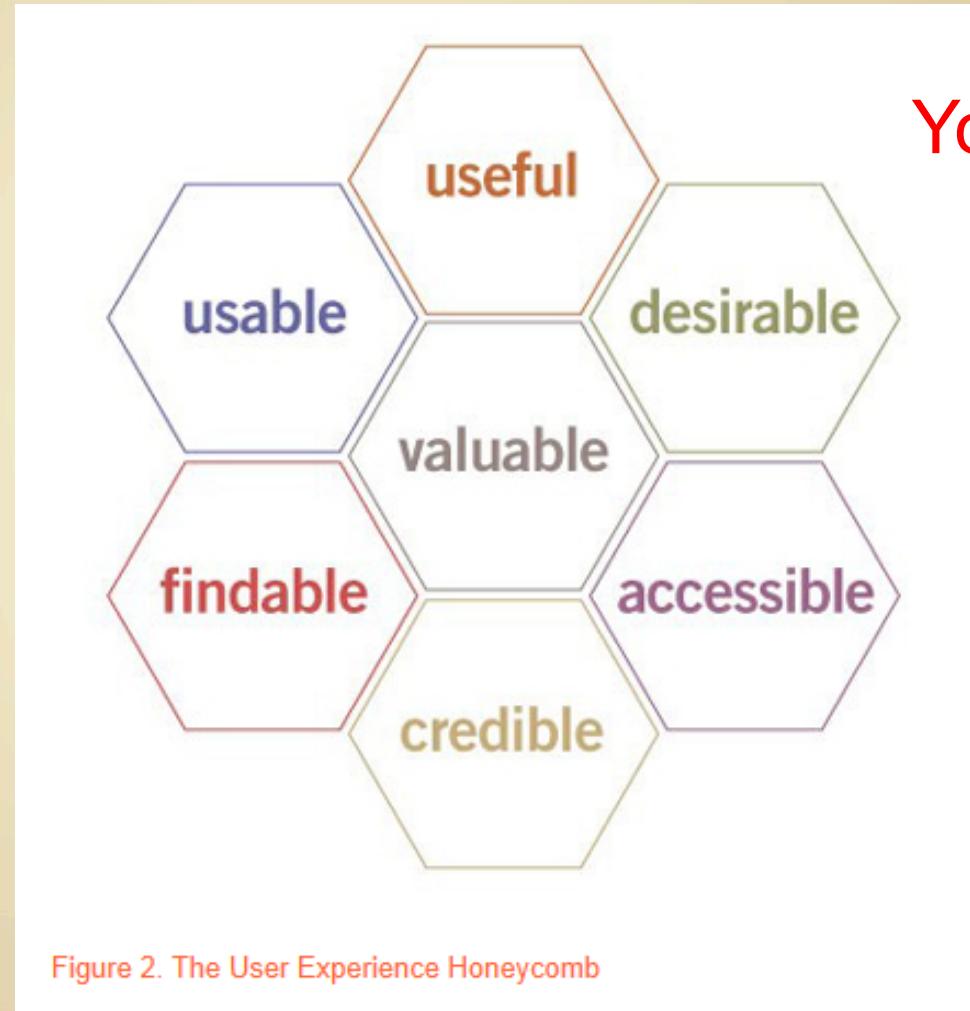


# What is user experience?

- “User experience encompasses all aspects of the end-user’s interaction with the company, its services, and its products. The first requirement for an exemplary user experience is to meet the **exact needs** of the customer... Next comes simplicity and elegance that produce products that are a joy to own, a joy to use. True user experience goes far **beyond** giving customers what they say they want, or providing checklist features...” (Nielsen-Norman Group)
- “User experience and interface design in the context of creating software represents an approach that puts the user, rather than the system, at the center of the process. This philosophy, called user-centered design, incorporates user concerns and advocacy from the beginning of the design process and dictates the needs of the user should be foremost in any design decisions.” (Microsoft)

# What makes good user experience?

Morville's honeycomb





## EVALUATION: WHY, WHAT, HOW AND WHEN?

“...If you don’t have user-testing as an integral part of your design process you are going to throw buckets of money down the drain.”

*- Bruce Tognazzini, Nielsen Norman Group*

# It's my fault! (or is it?)



- Users of a design often assume blame, believe they're 'stupid' or 'clumsy'
- The simpler or trivial a task *appears*, the more users tend to blame themselves
- “If I do some *action A* just prior to some *result R*, then I conclude that *A* must have caused *R*, even if there was no relationship between the two.”
- Result is a virtuous circle: silence, guilt, helplessness → designs that are frustrating and confusing.

# Evaluation is how you assess you got it right...

- “Evaluation” is closely related to the design issues and techniques.
- However carefully we follow HCI guidelines and standards, when designing systems we need to step back from the developing system and **evaluate** it. Otherwise we may create a monster ...

Dr Frankenstein carefully followed the  
“Guidelines and Standards for Creation of  
an artificial Human Being”



# ...but, evaluation is NOT testing!

- Evaluation is not the same as “testing”, because testing is designed to **find bugs** whereas evaluation has a different focus.
- A system may be totally bug free and yet still have a poor interface
- Evaluation is either predictive (does not involve users and is quick and relatively inexpensive) or usability testing (which does involve users).

# Why, what, how and when to evaluate...

Evaluation is a continuous process that examines:

- ❑ **Why:** to check users' requirements and that they can use the product and they like it
- ❑ **What:** a conceptual model, early (often paper) prototypes of a new system and later, more complete prototypes
- ❑ **How:** in natural and laboratory settings
- ❑ **When:** throughout design; finished products can be evaluated to collect information to inform new products

# When to evaluate...

- **Formative evaluations:** done during design and development to check that a product continues to meet users' needs. Results are then fed back into the design activity
- **Summative evaluations:** done to assess the success/quality of a finished product



National Institute of Standards and Technology (NIST), the International Standards Organization (ISO), the British Standards Institute (BSI)



# EVALUATION TYPES

# TYPES OF EVALUATION ARE...

- **Settings not involving users** - e.g. to predict, analyze and model aspects of the interface
- **Controlled settings** - involve users; e.g. usability testing and experiments in laboratories
- **Natural settings** - involve users; e.g. field studies and ‘in the wild’ studies to see how the product is used in real world

# ANY SETTINGS WITH NO USERS...

- ❑ Settings where the researcher has to **imagine** or **model** how an interface is likely to be used
- ❑ **Inspection methods** are commonly employed:
  - ✓ Heuristic evaluation
  - ✓ Cognitive walkthroughs
  - ✓ Analytics e.g. web analytics
  - ✓ Models e.g. Keystroke Level Model, Fitts' Law

# Heuristic evaluation

- A **heuristic** is a guideline, general principle or rule of thumb e.g. “Students who do all the tutorial exercises do better in examinations.” There may be exceptions, but on the whole it’s true.
- It is a **discount** usability inspection method and the focus is on the interface. It is undertaken by HCI experts.
- User-centred, highly practical approach, natural behaviour in environment or lab.

# Nielsen's 10 Usability Heuristics

1. Visibility of system status
  2. Match between system and the real world
  3. User control and freedom
  4. Consistency and standards
  5. Error prevention
  6. Recognition rather than recall
  7. Flexibility and efficiency of use
  8. Aesthetic and minimalist design
  9. Help users recognize, diagnose, and recover from errors
  10. Help and documentation
- How?

# Shneiderman's Eight Golden Rules

1. Strive for consistency.
2. Enable frequent users to use shortcuts.
3. Offer informative feedback.
4. Design dialog to yield closure.
5. Offer simple error handling.
6. Permit easy reversal of actions.
7. Support internal locus of control.
8. Reduce short-term memory load

# How to do a heuristic evaluation..

- *Briefing session:* Description of what the expert evaluators are supposed to do
- *Evaluation period:* 1-2 hour period during which evaluators independently inspect the product.
  - Usually two passes are done: first pass gives overall feel for flow of interface; second pass focuses on specific interface elements
  - Experts often choose specific tasks to focus evaluation
- *Debriefing session:* Experts come together to discuss their findings, to prioritize problems, and to suggest solutions

# Cognitive walkthroughs

- Focus on ease of learning.
- Designer presents an aspect of the design and usage scenarios.
- Expert is told the assumptions about user population, context of use, task details.
- One or more experts walk through the design prototype with the scenario.

# The 3 questions

Experts are guided by **3 questions**:

- Will the correct action be sufficiently evident to the user?
- Will the user notice that the correct action is available?
- Will the user associate and interpret the response from the action correctly?

As the experts work through the scenario they note problems.

# Controlled setting with users...

- ..enable evaluators to **control** what users do, when they do it, and for how long
- ..enable evaluators to **reduce** outside influences and distractions
- **Usability testing:** typical example of controlled evaluation
  - ✓ generally done (or used to) in laboratories
  - ✓ primary goal to determine if an interface is usable by the intended user population
  - ✓ can be supplemented by observation, interviews, experiments

# What is usability testing...

“Extent to which a product can be used by specified users to achieve specified goals with **effectiveness**, **efficiency** and **satisfaction** in a specified context of use.”

- ISO 9241

- ✓ **Effectiveness:** accuracy and completeness; does it do what users need?
- ✓ **Efficiency:** resources expended; how quickly can users perform tasks?
- ✓ **Satisfaction:** comfort and acceptability of system; how pleasant is it to use? (Koivunen and May, 2002; Nielsen, 2012)

# HOW DO I TEST USABILITY?

- ❑ Goals and questions focus on how well users **perform tasks** with the product
- ❑ Data collected by **video**, interaction **logging**, and **thinking-aloud**
- ❑ User satisfaction **questionnaires** and **interviews** provide data about users' opinions (subjective measures)
- ❑ **Comparison** of products or prototypes is common



Testing is central!

# What type of data should I collect (metrics)?

Focus is on **time to complete task AND number and type of errors** (Wixon and Wilson, 1997):

- Time to complete a task
- Time to complete a task after a specified time away from the product
- Number of users successfully completing a task
- Number and type of errors per task
- Number of errors per unit of time
- Number of times online help and manuals accessed
- Number of users making an error

# How many participants is enough?

- The number is a practical issue
- Depends on:
  - ✓ schedule for testing;
  - ✓ availability of participants;
  - ✓ cost of running tests
- Acceptable to have **5-12 users** (Dumas and Redish, 1999); Nielsen (2000) recommends **5 users**
- Some experts argue that testing should continue until no new insights are gained

# What happens in a Usability Lab...

Video cameras

Monitors

Microphones

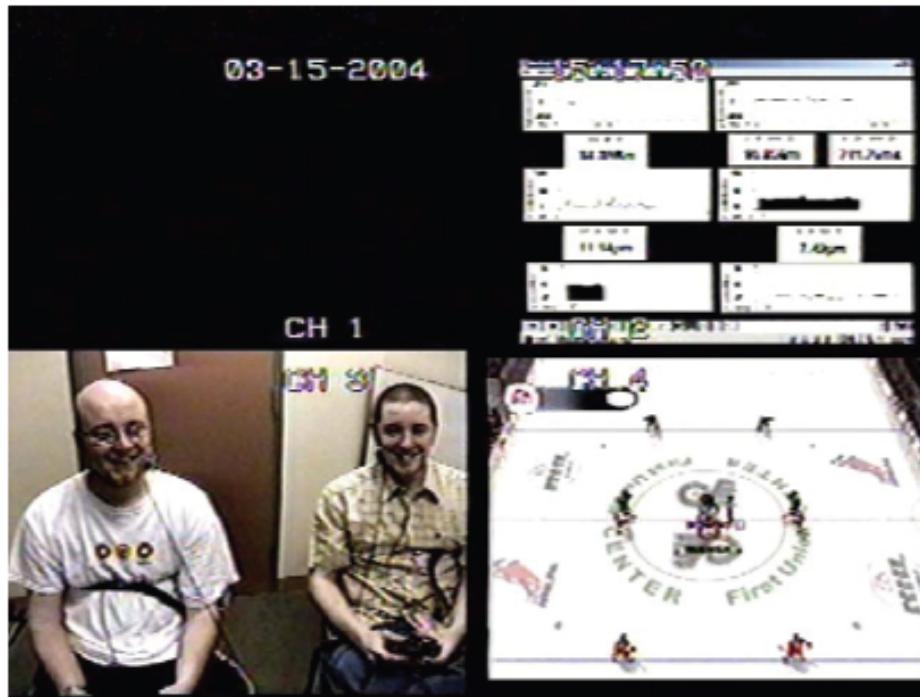
Mirror



**Figure 14.1** A usability laboratory in which evaluators watch participants on a monitor and through a one-way mirror

Adapted from Preece, Sharp and Rogers (2015)

# A COLLABORATIVE GAME (15 min)



What to measure?  
Kind of setting?  
  
How much control?  
Which methods?

**Figure 13.2** The display shows the physiological data (top right), two participants, and a screen of the game they played

Source: Mandryk and Inkpen (2004) Physiological Indicators for the Evaluation of Co-located Collaborative Play, CSCW'2004, pp. 102–111. ©2004 Association for Computing Machinery, Inc. Reprinted by permission.

# Remote usability testing



UserZoom

OptimalWorkshop

The screenshot shows a user interface for 'OptimalWorkshop'. At the top right are buttons for 'View Instructions', 'Leave a comment', and 'Finished'. On the left, there's a sidebar with a logo and a list of tasks:

- Purchase a cell phone online
- Discounted cell phones
- Assessments for my cell phone
- Downloadable games for my cell phone
- Find data that my next telephone and internet bill is due
- A tool to calculate the best cell phone plan for me
- What to do when my cell phone has been stolen or lost
- Other video surveillance at BananaCom

The main area is divided into sections:

- Internet:**
  - Internet, television & speed test
  - Deals for home internet and phone bundles
- Account:**
  - Pay my bill online
  - Change my home internet plan online
- Home Phone:**
  - Find new services for my home phone
  - International calling rates for my home phone
- Group:**
  - BananaCom's related phone numbers

# ..OR NATURAL SETTINGS WITH USERS

- **Field studies** to evaluate people in their natural settings
  - ✓ help identify opportunities for new technology
  - ✓ establish requirements for a new design
  - ✓ facilitate the introduction of technology, or inform deployment of existing technology in new contexts
- Goal is to be **unobtrusive** and not to affect what people do during evaluation
- **Methods** typically used are observation, interviews, and logging
- More recently **virtual** studies!

# A USABILITY TESTING CASE STUDY



# Usability testing the iPad!

- ❑ 7 participants with 3+ months experience with iPhones
- ❑ Signed an informed consent form explaining:
  - ✓ what the participant would be asked to do;
  - ✓ the length of time needed for the study;
  - ✓ a promise that the person's identity would not be disclosed; and
  - ✓ an agreement that the data collected would be confidential and would be available to only the evaluators
- ❑ Then they were asked to explore the iPad

# Examples of tasks...

App or website	Task
iBook	Download a free copy of <i>Alice's Adventures in Wonderland</i> and read through the first few pages.
Craigslist	Find some free mulch for your garden.
eBay	You want to buy a new iPad on eBay. Find one that you could buy from a reputable seller.
Time Magazine	Browse through the magazine and find the best pictures of the week.
Epicurious	You want to make an apple pie for tonight. Find a recipe and see what you need to buy in order to prepare it.
Kayak	You are planning a trip to Death Valley in May this year. Find a hotel located in the park or close to the park.

**Table 14.1** Examples of some of the tests used in the iPad evaluation (adapted from Budiu and Nielsen, 2010).

Source: Copyright Nielsen Norman Group, from report available at <http://www.nngroup.com/reports/>.

# Example of the equipment used...



**Figure 14.6** The setup used in the Chicago usability testing sessions

Source: Copyright Nielsen Norman Group, from report available at <http://www.nngroup.com/reports/>.

# Problems and actions...

## □ Problems detected:

- ✓ Accessing the Web was difficult
- ✓ Lack of affordance and feedback
- ✓ Getting lost
- ✓ Knowing where to tap

## □ Actions by evaluators:

- ✓ Reported to developers

## □ Accessibility for all users important



Adapted from Preece, Sharp and Rogers (2015)

# Developing an evaluation plan..

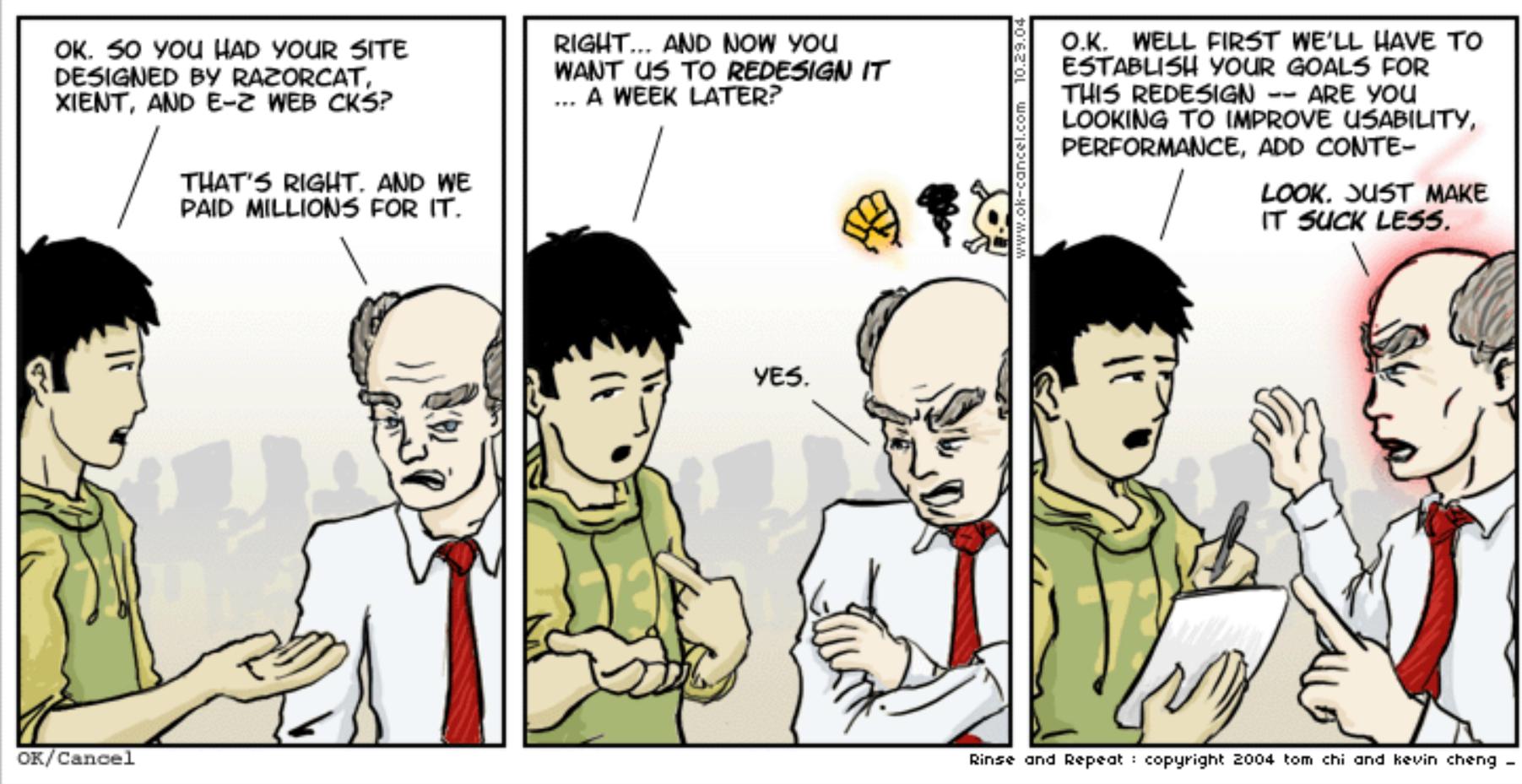
- Set evaluation (usability) goals
- Select the tools and techniques you will use
- Establish membership of the evaluation team

“Produce a set of heuristics derived from relevant standards and guidelines; choose the method and the evaluators; run the test (storyboard or prototype); collect the comments; analyse the results; identify areas where the design needs to be improved.”

# Some key points

- Evaluation and design are very closely integrated.
- Different evaluation methods are used for different purposes at different stages of the design process and in different contexts of use.
- Evaluators mix and modify methods to meet the demands of evaluating novel systems.
- Some of the data gathering methods (see lecture notes) are used in evaluation as for establishing requirements and identifying users' needs, e.g. observation, interviews, and questionnaires.
- Evaluations can be done in controlled settings such as laboratories, less controlled field settings, or where users are not present.

# We all need to evaluate!



# SOME USEFUL READING...

