Evaluating Interactions

Lecture 6

Adam Girard

Why Evaluate?

Determine how usable interaction is

Identify good & bad features to inform design

Compare design choices to assist in decisions

Observe effects of specific interfaces on users

Ensuring that product/interface is well designed

Things to Remember

Evaluation is key component of HCI

Evaluation is a process, not an event

Making things better starts by evaluation

Who to evaluate with?

Identifying Participants

- Random sampling
- Convenience sampling
- Sampling to saturation (hard to do)
- Make sure representative of user group
- How many is enough?
  - Depends on aims of evaluation & constraints

What to evaluate?

What is the goal of the evaluation?

- The effect of x on y? Gaining feedback on interface?
- Need to be clearly set out before evaluation

When to evaluate?

Formative

- Checking that designs meet user needs
- Dev. of sketches & prototypes
- Opportunistic evaluations

Summative

- Evaluating to see what needs improved
- Evaluating whether reached criteria

Where to evaluate?

In the wild

- In environment where tech is used

In the lab

- When looking for control & causal conclusions
- Can be built to emulate environment

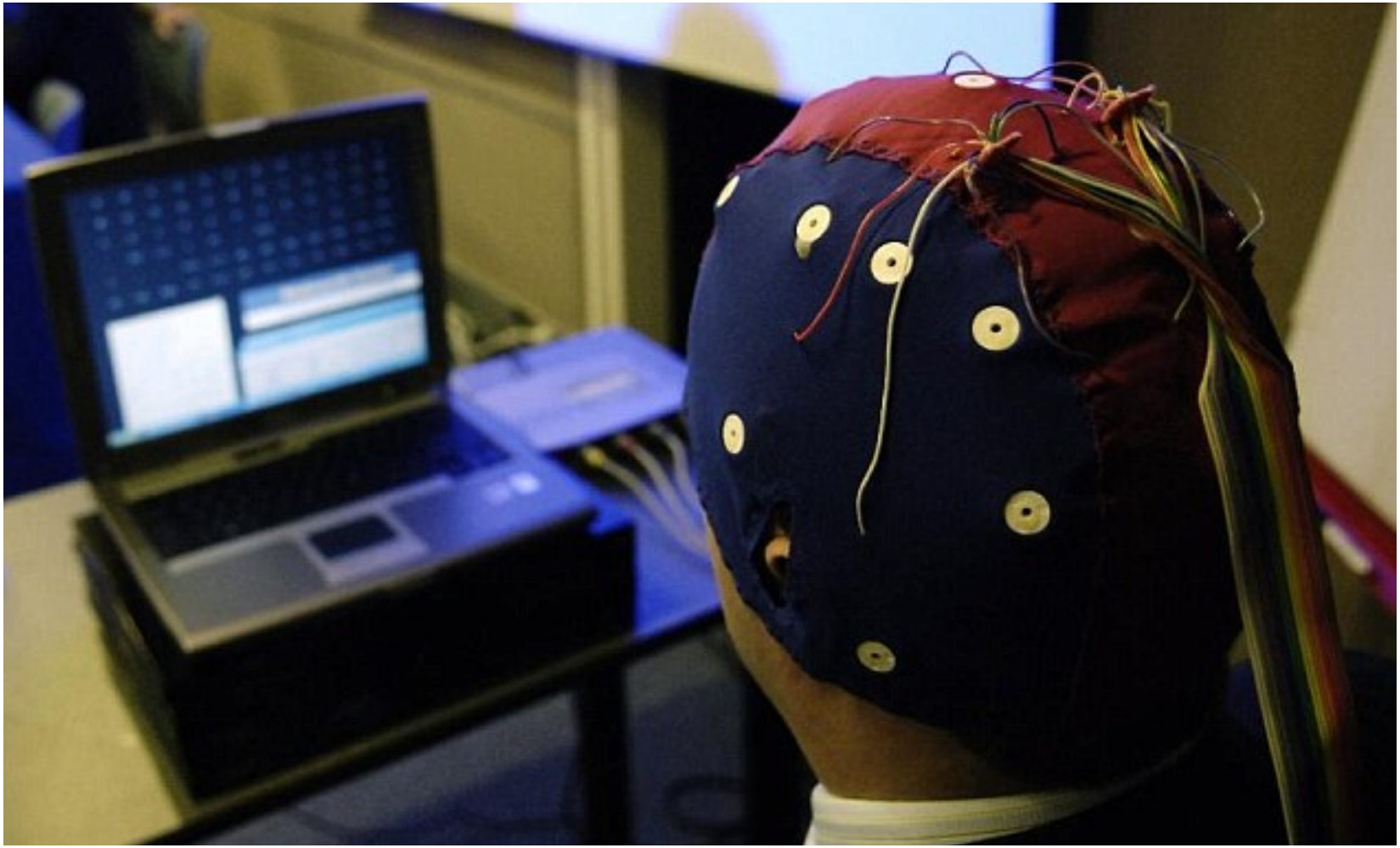
Remote testing

- Online questionnaires etc

Types of Evaluation

1. Controlled settings with users
   - Lab studies, living labs
2. Natural settings with users
   - Field studies, Ethnography
3. Not involving users
   - Heuristic evaluation, expert evaluation, inspection, walkthroughs, user modelling

Controlled Settings



Usability Testing

Testing tenets of usability

- Efficiency (time to complete task)
- Effectiveness (completion/error rate)
- Satisfaction (questionnaire, SUMI)

Focusing on specific:

- User
- Task
- Environment

Usability Testing

Mix of data sources

- Experiments
- Observation/Video of interaction
- Interviews
- Questionnaires

All in controlled setting

Used to inform design and test issues

Controlled Experiment

Scientific Method

Control is key

- Reduction of confounds/order effects

Aim to investigate hypotheses about how the designs affect DV

Pre defined task/goal

Controlled Experiment

Comparison of design solutions

Results can feedback into redesign

Typically termed *usability engineering*

Controlled Experiment

Independent variables (IV's)

- Variables controlled by the experimenter

Dependent variables (DV's)

- Variables being observed

Types of Experiment Design

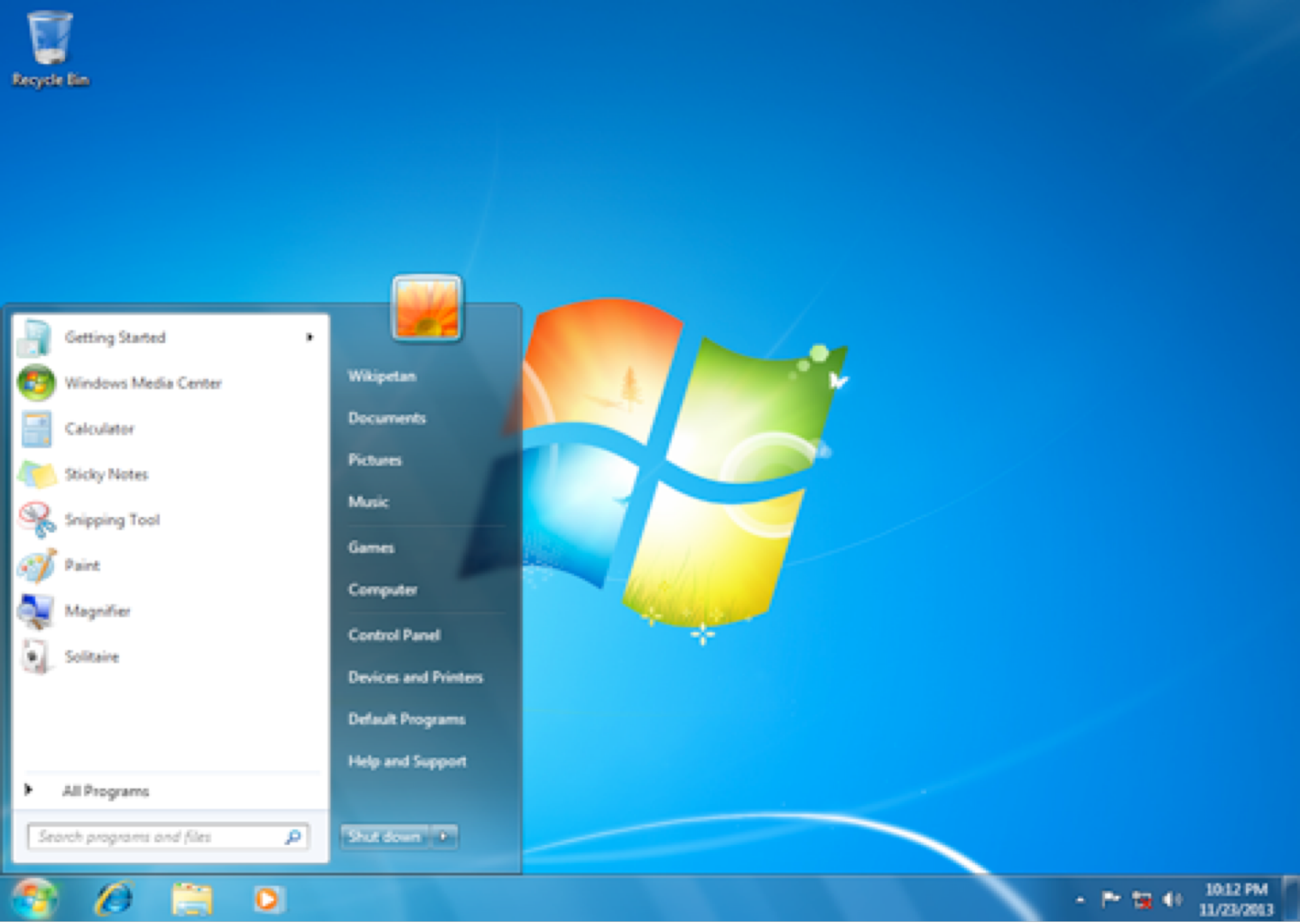
Between-participants

Within-participants

Benefits and drawbacks

- Practice effects (counterbalancing helps)
- Large N needed (more so in Between)

http://freehdwallpaperdownloads.com/windows/windows-8-desktop-hd-wallpaper.html/



http://upload.wikimedia.org/wikipedia/en/b/bd/Windows_7.png

What questions should you ask yourself?

What hypotheses could you explore?

What would the Dependent Variables, and Independent Variables be?

What do we need to control?

What are the pros and cons?

The ecological validity conundrum

Controlled experiments are very useful

- Causal inference
- Specificity of effect
- Replicable and robust

But are they realistic?

- Artificiality of scenario/lab environment
- Hawthorne effect

Increasing ecological validity in experiments

Use representative participants

Make the environment as realistic as possible

Make the tasks & scenario as realistic as possible

Natural Setting

Field Studies

Measuring & observing in “real world”

Quite messy/ no specific hypothesis

Range from few mins to months

Field Studies

Diary studies

- Participants given diaries to complete
- Asked specific questions in diaries

Mobile Experience Sampling

- Complete diary on phone
- Asked context related questions
  - eg. What are you doing in the kitchen?

Observations

Film/observe in situ

Best to have a framework

- List of aspects/questions you want to focus on

The Person- Who is using the tech
The Place- When are they using it
The Thing- What are they doing with it

Observations

Passive/ participant observer

The importance of writing up notes/experiences as close to events as possible

- Important for validity of observations

Stop when you stop learning new things

Ethnography

Observation of users in their natural environment e.g. where the product is used

Can lead to insight into

- Problems (amount and significance) in interaction
- Ideas for solutions

Ethnography

Examples of data collected

- Conversation & semi structured interviews
- Researcher observations & Q&A
- Descriptions of activities or environments
- Memos and notices in the environment
- User stories

Ethnography

Benefits

- High ecological validity
- Great for identifying how design fits into the "real world"

Drawbacks

- Lack of control in design
- Data can be tricky and cumbersome to analyse
  - Video, audio coding etc
- Fluidity of interpretation

In the Wild Research

Context driven enquiry

User's existing environment

Naturalistic & unobstrusive

Diary studies, life logging

In the Wild- Pitfalls

No control

- What caused the behaviour?

Relying on participant/observer a lot

Interpretation of the data coder

Are the behaviours representative?

Considerations

Ethics

- Recording permission, length of time, anomymity/confidentiality

What if the tech/eval equipment breaks?

- Can the user fix it?

Not involving users

Useful When…

Hard to get hold of users

Expensive to involve users

Takes too long to involve users

Heuristic Evaluation

Method for finding usability problems

Popularised by Jakob Nielsen

“Discount” usability engineering

Used as tool to categorize problems

Heuristic Evaluation

Systematic inspection to see if interface complies to guidelines

Method

- 3-5 inspectors
- usability engineers, end users, double experts…
- inspect interface in isolation (~1–2 hours for simple interfaces)

compare notes afterwards

Works for paper, prototypes, and working systems

Points of Variation

Evaluators

Heuristics used

Method employed during inspection

Evaluators

These people can be novices or experts

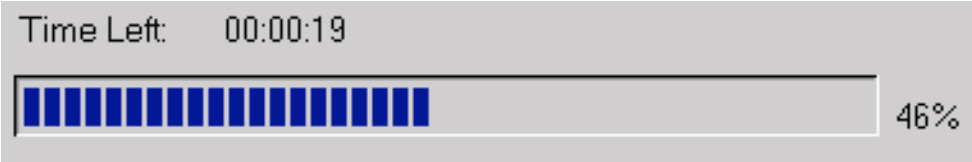
- "novice evaluators"
- "regular specialists"
- "double specialists" *(- Nielsen)*

Each evaluator finds different problems

Best evaluators find hard & easy problems

Nielsen's Heuristics

Visibility of system status



Match between system & real world

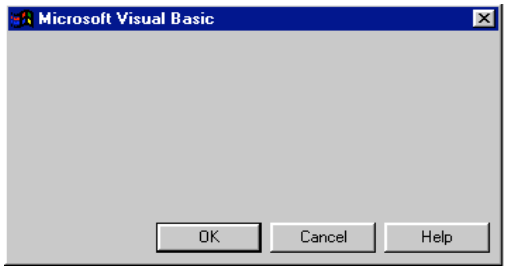
- Poor error messages fail this test

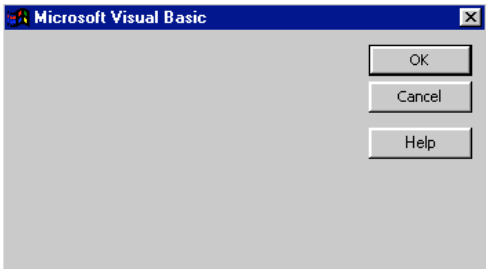
User control and freedom

- Undo, redo, easy error correction
- Not forced into particular path

Nielsen's Heuristics

Consistency & standards





Error prevention

Recognition rather than recall

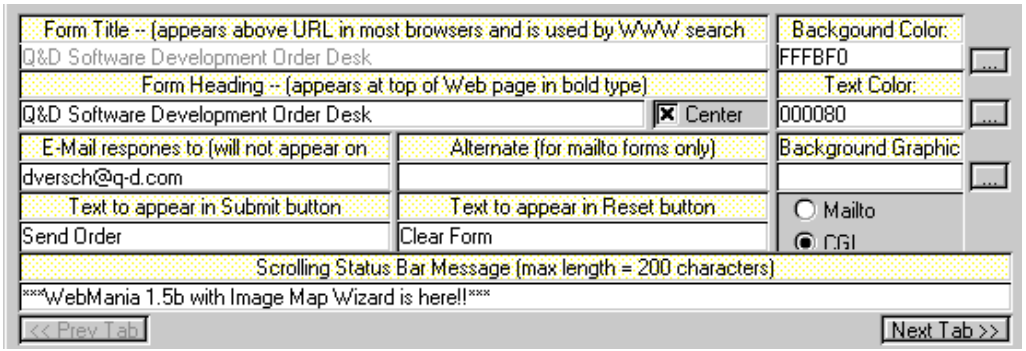
- Violated by large ATS menus

Nielsen's Heuristics

Flexibility & efficiency of use

- Allow shortcut keys
- Multiple methods for all types of users

Minimalist design



Nielsen's Heuristics

Help users recognise, diagnose and recover from errors

Help & documentation

Phases of a heuristic evaluation

1. Pre-evaluation training
   - give evaluators domain knowledge & information on the scenario
   - Use script

2. Evaluate interface independently
   - Expert takes 1-2 hours independently inspecting product
   - Tend to do 2 passes through interface
   - Experts give themselves tasks

Phases of Heuristic Evaluation

3. Rate each problem for severity

4. Aggregate results

5. Debrief
   - Come together to discuss & prioritise issues

Severity ratings

Each evaluator rates individually:

- 0 - don't agree that this is a usability problem
- 1 - cosmetic problem
- 2 - minor usability problem
- 3 - major usability problem; important to fix
- 4 - usability catastrophe; imperative to fix

Consider both impact and frequency

Single inspector

Average over six case studies

- 35% of all usability problems;
- 42% of the major problems
- 32% of the minor problems

Not great, but much better than finding no problems with no evaluators!

Single inspector

Varies according to

- difficulty of the interface being evaluated
- the expertise of the inspectors

Average problems found by:

- novice evaluators - no usability expertise - 22%
- regular specialists - expertise in usability - 41%
- double specialists - experience in both usability and the particular kind of interface being evaluated – 60%

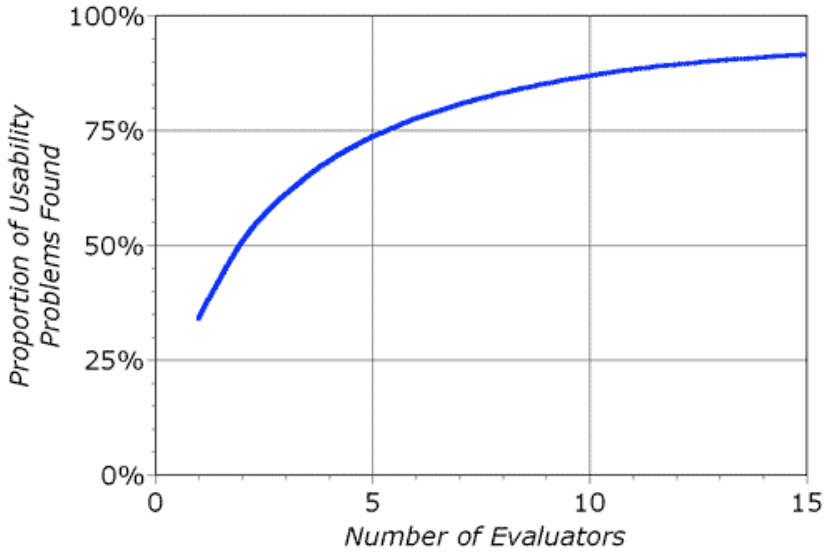
Tradeoff

- novices poorer, but cheaper!

Multiple evaluators

3-5 evaluators find:

- 66-75% of usability problems
- different people find different usability problems
- only modest overlap



Individuals vs. teams

Nielsen's Recommendations

- recommends individual evaluators inspect the interface alone

Why?

Benefits of solo interface inspection

- evaluation is not influenced by others
- independent and unbiased
- greater variability in the kinds of errors found
- no overhead required to organize group meetings

Self Guided vs. Scenario Exploration

Self-guided

- open-ended exploration
- Not necessarily task-directed
- good for exploring diverse aspects of the interface, and to follow potential pitfalls

Self Guided vs. Scenario Exploration

Scenario Based Inspection

- step through interface using representative tasks
- ensures problems identified in relevant portions
- ensures that features of interest are evaluated
- but limits the scope

How useful are inspection methods?

Discount methods useful for practitioners

Not a replacement for user testing!

They are not rigorous scientific methods

- All inspection methods are subjective
- No inspection method can compensate for inexperience or poor judgement

Cognitive Walkthroughs

Simulate user problem solving process

Checks to see how user would progress

Look at ease of learning

How to do a Walkthrough

1. Define characteristics & tasks of typical user, description of interface & clear actions needed to complete task

2. Designer & expert evaluator do the analysis

How to do a Walkthrough

3. Evaluator walks through actions sequences for each task placing in context of typical scenario

- Will correct action be evident to user?
- Will user notice correct action available?
- Will user associate & interpret response from action correctly?

How to do a Walkthrough

4. Record:
   - Assumptions about what would cause problems & why
   - Notes on side issues & design changes
   - Compile summary of results

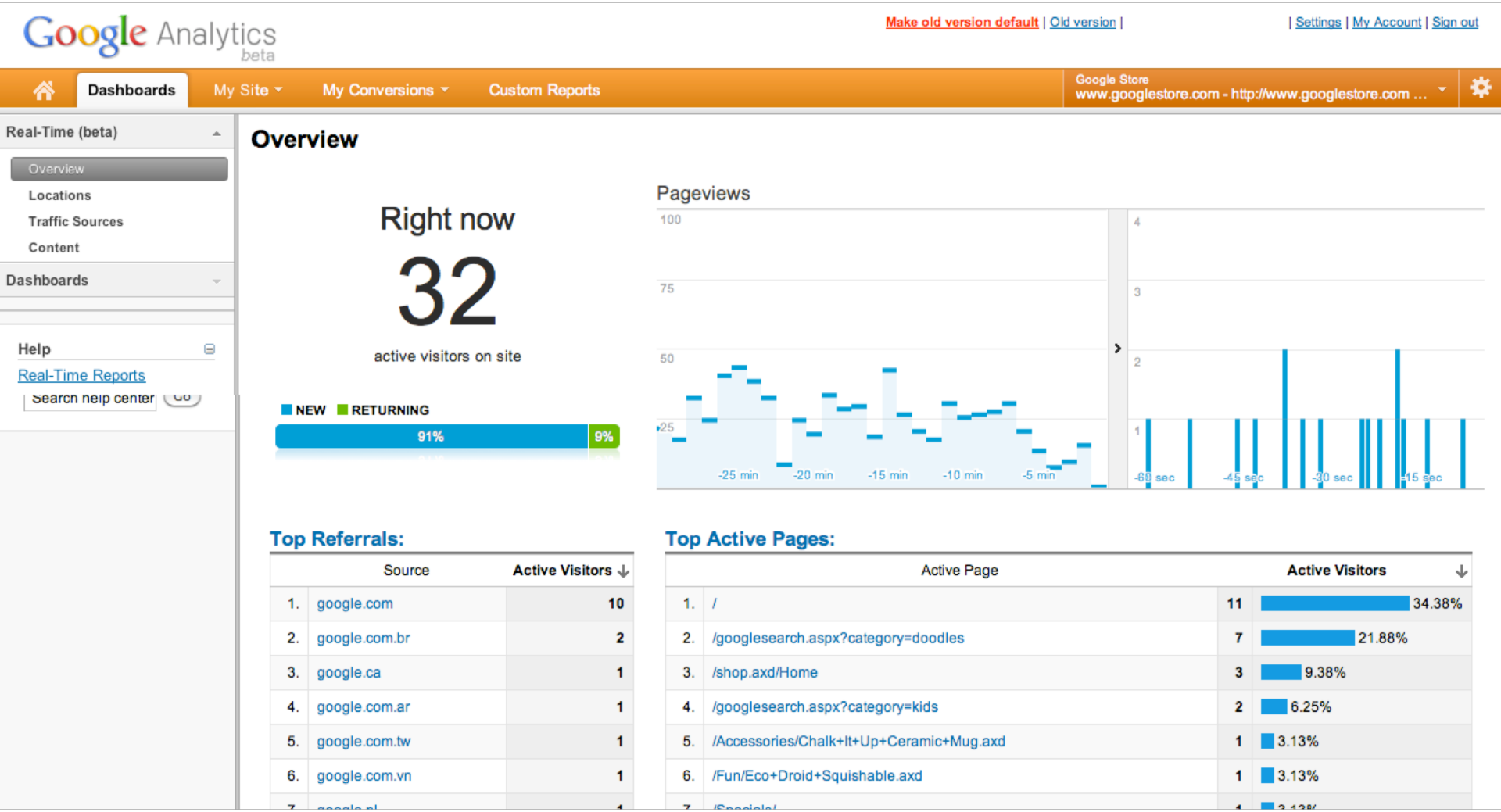
5. Revise design to fix problem

Analytics

Log data about events

Useful for identifying use/uptake

Tell you what but not why



http://analytics.blogspot.ie/2011/09/whats-happening-on-your-site-right-now.html

Predictive Models

Use formulas or steps to derive outcomes

Estimating efficiency of systems

GOMS Model

Goals- state that user wants to achieve

Operators- Cog processes & physical actions needing to be performed

Methods- learned procedures for accomplishing goal

Selection Rules- used to determine which method when more than one available

Keystroke Level Model

Provides numerical predictions of user performance

Standard set of times for actions (key presses, click mouse, click button, decision, system response rate)

Predicted time calculated by summing these aspects

Benefits & Drawbacks

Compare different interfaces/tasks easily

Not commonly used for evaluation

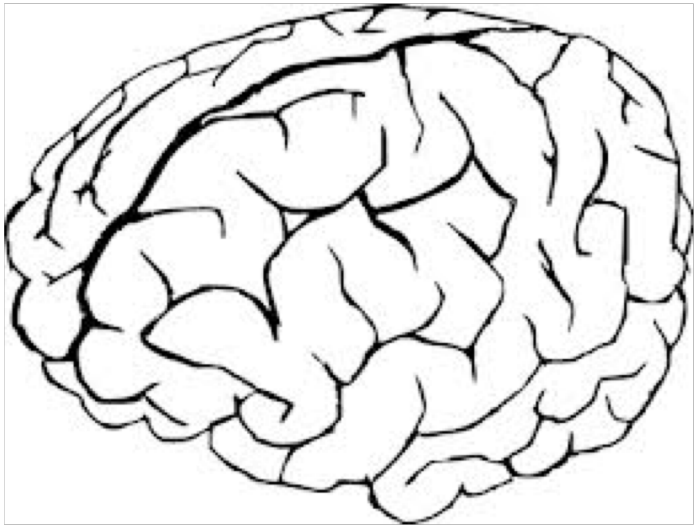
- Limited scope to data entry tasks
- Can't really model for errors

Doesn't take into account individual differences, multitasking, interruptions

Methods:
Tools for Gathering Data

Think Aloud Studies

People interact and talk whilst conducting task



Exploratory or specific

The sound of silence

- Stop-start nature of this method

Interviews

Conversation with a purpose

4 main types

- Unstructured
- Semi Structured
- Structured
- Focus Group

What interview type to use?

Depends on:

- Specificity of goals
- Purpose of the interview
- Stage in UCD/ID cycle

Unstructured interviews

Exploratory

Talk around an area

Planning the areas for discussion

Open questions

- What do you feel are the advantages of x?

Can explore topics as they come up

Structured Interviews

Predetermined questions

Standardised

Closed questions

- Which of the following do you visit frequently (then list websites)

Semi Structured Interviews

Basic script used with all participants

Mix of Structured and Unstructured Interview

Tips on Interviews

Short questions

- Break down complex questions

Explain or try to avoid jargon

Keep questions neutral

- Not "Why did you like the e-reader?" But "Is there anything you liked about the e-reader?"

Tips on Interviews

Do not respond with bias

Sequence your interview logically

- General, then specific

Enrich the experience

- Screenshots, physical tech

Interviews

Consistency

- Unstructured vs Structured

Time to analyse data

Richness and depth of data

Questionnaires

Collecting demographic and opinion data

No researcher present (mostly)

- Need to be clearly worded

Need to think of purpose of questionnaire

Example- Demographic Questionnaire

Demographic Questionnaire

1. Gender:
   - ▫ Male
   - ▫ Female

2. Age: ……..yrs …….months

3. What year of your Psychology Undergraduate degree are you in at the moment?
   - ▫ Y2
   - ▫ Y3
   - ▫ Y4

4. Are you an honours year psychology student?
   - ▫ Yes
   - ▫ No

5. Have you in the past or are you currently enrolled in any of the Differential Psychology courses offered by the Psychology Department?
   - ▫ Yes
   - ▫ No

6. If so, which are you enrolled in or have you completed?

   Taught Courses
   - ▫ Differential Psychology (Y3-Term1)
   - ▫ Basic Tendencies of Personality (Y4-Term1)
   - ▫ Causes and Consequences of Personality (Y4- Term1)

   Project Choices
   - ▫ Differential Psychology Research Project (Y3)
   - ▫ Dissertation in Personality and Genetics (Y4)

Example- Likert Scale Questionnaire

Please place a tick (✓) in the box showing your level of agreement with each of the statements below.

		Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
826	I thought the interaction with my partner was complicated					
804	I liked the interaction					
802	When interacting I didn't always know what to do next					
803	I felt in control when interacting with my partner					
831	I felt the interaction needed improvement					
817	I would not interact with this partner again					
822	I found the interaction with my partner confusing					

Questionnaires

Scales

- Likert (3 point, 5 point, 7 point)
- E.g Strongly Disagree- Strongly Agree
- Semantic Differential

Acquiescence Response bias

- Balance and vary negative and positive items

The importance of “Don’t know”

Tips on Questionnaires

Think of what/who you need to ask

- Online or offline administration?

Give clear instructions

Be efficient

What information are you wanting?

- Demographic
- Scale variable (for comparison etc)

Designing your own attitude questionnaire

Is time consuming!

Iterative

- Checked for face validity with group
- Piloted
- Reworded

Reliability and validity

Summative or Factor assumption

Which should I use?

The goal

The design stage

The participants

- E.g. Location

The resources

- Time
- Money
- Previous experience