

EECS 4441 Human-Computer Interaction

Topic #5: Evaluation – Part I

I. Scott MacKenzie

York University, Canada



Evaluation

- Test the usability and functionality of a system
- Occurs in a laboratory, in the field, and/or in collaboration with users
- Evaluates both design and implementation
- Should be considered at all stages in the design life cycle

Goals of Evaluation

- Assess extent of system functionality
- Assess effect of interface on user
- Identify specific problems

Topics – Evaluating Design

- Cognitive Walkthrough
- Heuristic Evaluation
- Review-based Evaluation



No user participation

Cognitive Walkthrough (1)

- Proposed by Polson et al.¹
 - Evaluates design on how well it supports users in learning tasks
 - Usually performed by expert in cognitive psychology
 - Expert “walks through” design to identify potential problems using psychological principles
 - Forms used to guide analysis

¹Polson, P., Lewis, C., Rieman, J., and Wharton, C., Cognitive walkthroughs: A method for theory-based evaluation of user interfaces, *International Journal of Man-Machine Studies*, 36, 1992, 741-773.

Cognitive Walkthrough (2)

- For each task walkthrough considers
 - What impact will interaction have on user?
 - What cognitive processes are required?
 - What learning problems may occur?
- Analysis focuses on goals and knowledge: Does the design lead the user to generate the correct goals?

Heuristic Evaluation

- Proposed by Nielsen and Molich¹
- Usability criteria (heuristics) are identified
- Design examined by experts to see if these are violated
- Example heuristics
 - System behaviour is predictable
 - System behaviour is consistent
 - Feedback is provided
- Heuristic evaluation “debugs” design

¹ Nielsen, J. and Molich, R., Heuristic evaluation of user interfaces, *Proceedings of CHI '90*, (New York: ACM, 1990), 249-256.

Review-based Evaluation

- Results from the literature used to support or refute parts of design
- Care needed to ensure results are transferable to new design
- Cognitive models used to filter design options; e.g., GOMS prediction of user performance
- Design rationale can also provide useful evaluation information



Evaluating Through User Participation

Laboratory Studies

- Advantages:
 - Controlled environment (high in *precision*)
 - Specialised equipment available
 - Data tend to be quantitative (not qualitative)
- Disadvantages:
 - Lack of context (low in *relevance*)
 - Difficult to observe several users cooperating
- Appropriate...
 - If system location is dangerous or impractical for constrained single user systems to allow controlled manipulation of use
 - To test research ideas

Field Studies

- Advantages:
 - Natural environment (high in *relevance*)
 - Context retained (though observation may alter it)
 - Longitudinal studies possible
- Disadvantages:
 - Lack control (low in *precision*)
 - Distractions, Noise, Chaos!
 - Labour intensive
 - Data tend to be qualitative (not quantitative)
- Appropriate
 - Where context is crucial for longitudinal studies

Topic: Evaluating Implementations

- Requires an artifact, such as
 - Simulation
 - Prototype
 - Full implementation
- Exception:
 - Wizard of Oz method (implementation is faked)

Experimental Evaluation

- Controlled evaluation of specific aspects of interactive behaviour
- Evaluator chooses hypothesis to be tested
- A number of experimental conditions are considered which differ only in the level of a manipulated variable (aka *independent variable*)
- Changes in behavioural measures (aka *dependent variables*) are attributed to different conditions

Experimental Components

- Subjects (today "Participants")
 - Who – representative
 - Include sufficient sample (as per related research)
 - State how participants were selected (random sampling preferred, but rarely done)
- Variables
 - Things to modify and measure
- Hypothesis
 - What you'd like to show
- Experimental design
 - How you are going to do it

Variables

- Independent variable (IV)
 - Circumstance changed to produce different conditions
 - E.g., interface style, number of menu items
- Dependent variable (DV)
 - Human behaviour measured in the experiment
 - E.g., time taken, number of errors, etc.

Hypothesis

- Prediction of outcome
 - Framed in terms of IV and DV
 - E.g., "error rate will increase as font size decreases"
- Null hypothesis:
 - States no difference between conditions
 - Aim is to disprove this
 - E.g. NH = "no change in error rate with font size"
 - Null hypothesis must be testable (i.e., "Interface A is better than interface B" is not testable)

Assign Test Conditions to Participants

- Within-subjects design
 - Aka “repeated measures design”
 - Each participant performs experiment under each condition
 - Transfer of learning possible
 - Less costly and less likely to suffer from user variation
- Between-subjects design
 - Each participant performs under only one condition
 - No transfer of learning
 - More users required
 - Variation can bias results

Analysis of Data

- Before you do any statistics:
 - Look at data (there may be outliers - wildly deviant measures)
 - Save original data
- Choice of statistical technique depends on
 - Type of data
 - Information required
- Type of data
 - Discrete - finite number of values
 - Continuous - any value

Analysis - Types of Tests

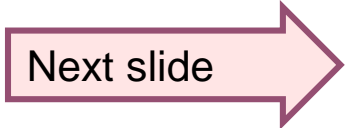
- Parametric
 - Assume normal distribution
 - Robust
 - Powerful
- Non-parametric
 - Do not assume normal distribution
 - Less powerful
 - More reliable
- Contingency table
 - Classify data by discrete attributes
 - Count number of data items in each group

Analysis of Data (continued)

- What information is required?
 1. Is there a difference?
 2. How big is the difference?
 3. How accurate is the estimate?
- Parametric and non-parametric tests mainly address point #1 above

User Study Example

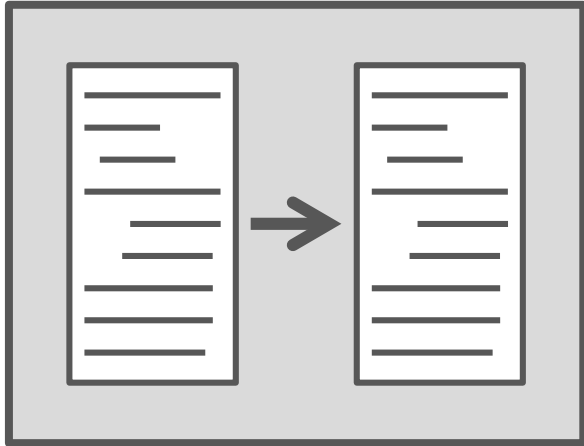
- Topic
 - Evaluating Icon Designs
- Source
 - Dix, A., Finlay, J., Abowd, G., & Beale, R. (2004). *Human-computer interaction* (3rd ed.). London: Prentice Hall, pp. 335-339.
- Research idea
 - It might be easier to remember the meaning of icons depending on how they are designed. Two designs of interest are "natural images" (based on a paper document metaphor) and "abstract images"



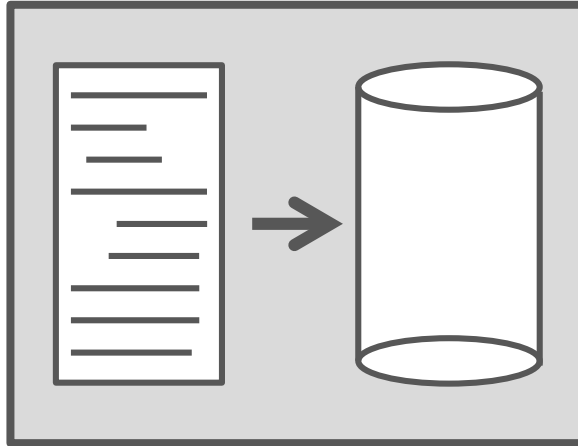
Next slide

Natural

(based on paper document metaphor)



Copy

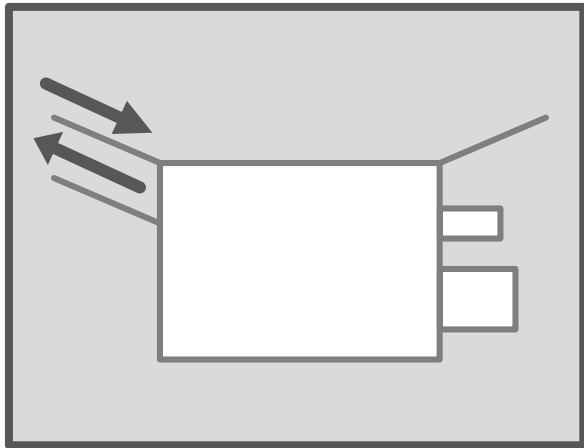


Save



Delete

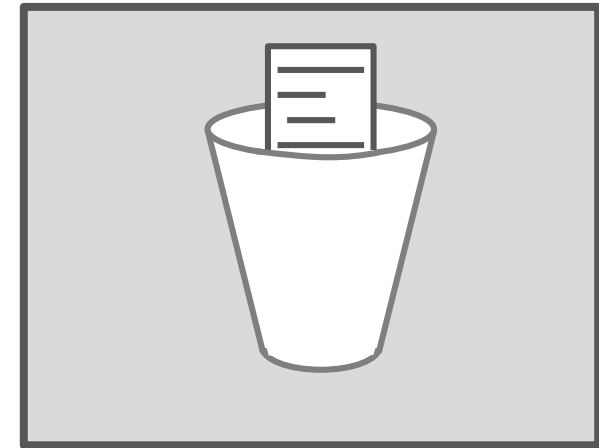
Abstract



Copy



Save



Delete

- Research question (hypothesis)
 - Will users remember natural icons more easily than abstract icons?
- Null hypothesis
 - There will be no difference between recall of the icon types
- Critique
 - Both the research question and the null hypothesis above are poorly formed because they are **not testable**
- A better formulation of the null hypothesis is...
 - The time to select the appropriate icon in response to a prompt is the same for natural icons and abstract icons

Writing Style and Terminology

- Be consistent!
 - In the Dix et al. text, icons designed according to a paper document metaphor are referred to in some places as "natural" and in other places as "concrete".
 - This is bad
 - Choose an appropriate term and stick with it!
- Similarly, is the study about “Icon Design” or “Icon Type”? (Both terms are used.)

Experiment Design

- Participants (information from Dix et al.)
 - 10
 - Demographics? ("sufficient participants from the intended user group")
 - Relevant experience? (no information given)
 - How selected, were they paid, etc.? (no information given)

Experiment Design (2)

- Apparatus
 - Not described
 - Were the tasks administered online or using a paper facsimile of the icons with responses entered on a sheet and timed by hand?

Experiment Design (3)

- Procedure

- Participants given a fixed amount of time to study the icons, then they are given a recall test
- How many icons were they required to identify?
- More details must be provided!
- Exposure to conditions counterbalanced with five participants per group:
 - AN group - Abstract first, Natural second
 - NA group - reverse order

Experiment Design (4)

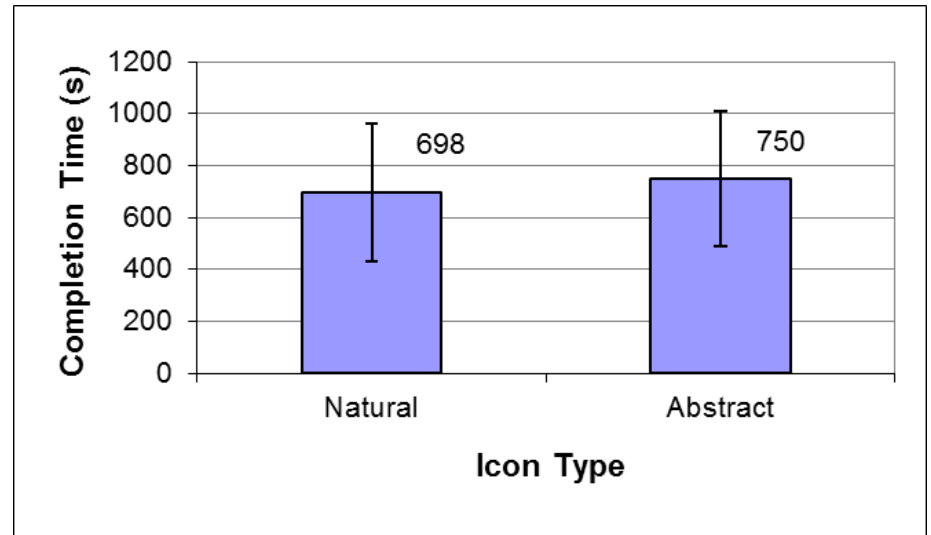
- Within-subjects
- Independent variable (aka factor)
 - Icon Type (levels: Natural, Abstract)
- Dependent variables
 - Task completion time (units: seconds)
 - Error rate (percentage of icons incorrectly identified)
- There is also a "Group" factor, which is between-subjects
 - 5 participants in AN group
 - 5 participants in NA group

• Results and Discussion



Excel

Completion Time (s) by Icon Type			
Participant	Icon Type		Group
	Natural	Abstract	
1	656	702	AN
2	259	339	AN
3	612	658	AN
4	609	645	AN
5	1049	1129	AN
6	1135	1179	NA
7	542	604	NA
8	495	551	NA
9	905	893	NA
10	715	803	NA
<i>mean</i>	697.70	750.30	
<i>SD</i>	265.13	258.75	
<i>Grand mean</i>	724.00		
<i>Percent longer</i>		7.5%	



ANOVA table for Completion Time (s)					
Effect	df	SS	MS	F	p
Group	1	67744.800	67744.800	0.466	0.5142
Participant (Group)	8	1163747.200	145468.400		
Icon Type	1	13833.800	13833.800	30.680	4.0E-4
Icon Type x Group	1	125.000	125.000	0.277	0.6128
Icon Type x P(Grou	8	3607.200	450.900		

Anova2

- Results and Discussion (2)
 - A partial write-up might be...

RESULTS AND DISCUSSION

Task Completion Time

The overall mean task completion time for the identification of icons was 724 s. The mean task completion time was lower for the Natural icons at 698 s. Abstract icons took about 7.0% longer to identify, with a mean of 750 s (see Figure 1). The difference was statistically significant ($F_{1,8} = 30.68, p < .001$). The Group effect, representing the order of presenting the two Icon Types to participants, was not significant ($F_{1,8} = 0.466, ns$). Thus, counterbalancing the order of presentation had the desired effect of cancelling any learning effect. There was also a non-significant Group by Icon Type interaction effect ($F_{1,8} = 0.277, ns$), suggesting an absence of asymmetric skill transfer.

*** Figure 1 about here ***

[discuss the results]

Error Rates

[present results on error rates]

Etc.

Thank You