01076568 Human Computer Interaction

Chapter 8 : Evaluation techniques

ดร.ชมพูนุท จินจาคาม
[kjchompo@gmail.com]

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

Outline

- What is evaluation?
- Goals of evaluation
- Evaluation through expert analysis
- Evaluation through user participation
- Choosing an evaluation method
- Consent form
- Summary

What is evaluation?

- Evaluation role is to access designs and test systems to ensure that they actually behave as we expect and meet user requirements.
- Ideally, evaluation should occur throughout the design life cycle, with the results of the evaluation feeding back into modifications to the design.

Goal of evaluation

- Evaluation has three main goals:
  - To assess the extent and accessibility of the system's functionality
  - To assess users' experience of the interaction
  - And to indentify any specific problems with the system.

- The system's functionality is important must accord with the user's requirements.
  - Evaluation at this level may measuring the user's performance with the system to assess the effectiveness of the system in supporting the task.
- User's experience of the interaction
  - How easy the system is to learn, its usability and the user satisfaction with it.
  - It may include his enjoyment and emotional response, particularly in the case of aim to entertainment.

- Identify specific problem with the design
  - This may aspects of the design which, when used in their intended context, cause unexpected results, or confusion amongst user.
- We will consider evaluation techniques under two broad headings: expert analysis and user participation.

Evaluation through expert analysis

- Cognitive walkthrough
  - Originally proposed by Polson and colleagues as an attempt to introduce psychological theory into the informal and subjective walkthrough technique.
  - The main focus is to establish how easy a system is to learn (by hands on, not by training or user's manual).

To do walkthrough, you need four things:

1. A specification or prototype of the system.
2. A description of the task the user is to perform on the system.
3. A complete, written list of the actions needed to complete the task with the proposed system.
4. An indication of who the users are and what kind of experience and knowledge the evaluators can assume about them.

The evaluation try to answer the following four questions:

1. Is the effect of the action the same as the user's goal at that point?
2. Will users see that the action is available?
3. Once users have found the correct action, will they know it is the one they need?
4. After the action is taken, will users understand the feedback they get?

- Heuristic evaluation
  - Heuristic is a guideline or general principle or rule of thumb that can guide a design decision or be used to critique a decision that has already been made. 3-5 evaluators is sufficient.
  - Severity rating on a scale of 0-4 (less-most)
    - 0 = I don't agree that this is a usability problem at all
    - 1 = Cosmetic problem only: need not be fixed unless extra time is available on project
    - 2= Minor usability problem: fixing this should be given low priority
    - 3= Major usability problem: important to fix, so should be given high priority
    - 4= Usability catastrophe: imperative to fix this before product can be released (Nielsen)

Nielsen's ten heuristics are:

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 an efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose and recover from errors
10. Help and documentation

- Model-based evaluation
  - Dialog models can be used to evaluate dialog sequences for problems, ex. Unreachable states, circular dialogs and complexity.
- Using previous studies in evaluation
  - Ex. Usability of different menu types, the recall command names, and the choice of icons.

Evaluation through user participation

- Styles of evaluation
  - Laboratory studies; take part in controlled tests.
  - Field studies; into the user's work environment in order to observe the system in action.

- Empirical methods: experimental evaluation
  - participants should be chosen to match the expected user population as closely as possible. And sample size must be large enough to be representative of the population.

- Empirical methods: experimental evaluation
  - Variables :
    - ตัวแปรต้น (สิ่งที่จะทำการทดลอง กำหนดขึ้นเพื่อทดสอบสมมติฐาน)
    - ตัวแปรตาม (ผลที่เกิดจากตัวแปรต้น เป็นตัวแปรที่ต้องทำการวัดค่า บันทึกผล)
    - ตัวแปรควบคุม (ตัวแปรที่ส่งผลต่อการทดลองให้คลาดเคลื่อนได้ จึงต้องควบคุมให้เหมือนกัน)

  - Hypothesis is the prediction of the outcome of an experiment.
    - Stating that a variation in the independent variable will cause a difference in the dependent variable.
    - By disproving the null hypothesis, which states that there is no difference in the dependent variable between the levels of the independent variable.

- Empirical methods: experimental evaluation
  - Experimental design
    - How many participants are available and are they representative of the user group?
    - Experimental method:
      - Between-subjects => participant is assigned to a different condition (experimental and control conditions)

      [Adv. learning effect resulting, Disadv. Require greater number of participants.]
      - Within-subjects => each user performs under each different condition.

      [This design can suffer from transfer of learning effect. Tackle by set group A(1-2) then group B(2-1)]

- Empirical methods: experimental evaluation
  - Statistical measures => look at the data (freak event) and save the data.

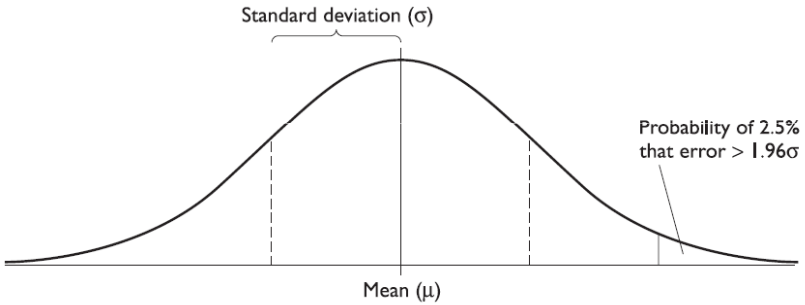

Figure 9.2 Histogram of normally distributed errors

Discrete = levels, numbers Ex. Screen color as red, green, blue
Continuour = any value Ex. Person's height, time taken to complete task

Table 9.1 Choosing a statistical technique

Independent variable	Dependent variable	
Parametric	- Assume data has come from probability distribution.	
Two valued	Normal	Student's *t* test on difference of means
Discrete	Normal	ANOVA (ANalysis Of VAriance)
Continuous	Normal	Linear (or non-linear) regression factor analysis
Non-parametric	- Fewer assumptions Ex. Movie rank order (1-4 stars)	
Two valued	Continuous	Wilcoxon (or Mann–Whitney) rank-sum test
Discrete	Continuous	Rank-sum versions of ANOVA
Continuous	Continuous	Spearman's rank correlation
Contingency tests	-True or false Ex. It's raining or it isn't raining, button is red, there're 3 menu	
Two valued	Discrete	No special test, see next entry
Discrete	Discrete	Contingency table and chi-squared test
Continuous	Discrete	(Rare) Group independent variable and then as above

An extensive and accurate analysis might ask about the data as:

- Is there a difference?

  Ex. Using Hypothesis testing. Answers are not yes/no but as 'we are 99% certain that selection from the menus of five items is faster than that from menus of seven items'.

- How big is the difference?

  Ex. 'Selection from five items is 260 ms faster than from the seven items'.

- How accurate is the estimate?

  Ex. 'Selection is faster by 260 ± 30 ms'. Or 'we are 95% certain that the difference in response time is between 230 and 290 ms'.

Example of non-parametric statistics

We will not see an example of the use of non-parametric statistics later, so we will go through a small example here. Imagine we had the following data for response times under two conditions:

condition A: 33, 42, 25, 79, 52
condition B: 87, 65, 92, 93, 91, 55

We gather the data together and sort them into order: 25, 33, 42, . . . , 92, 93. We then substitute for each value its rank in the list: 25 becomes 1, 33 becomes 2, etc. The transformed data are then

condition A: 2, 3, 1, 7, 4
condition B: 8, 6, 10, 11, 9, 5

Tests are then carried out on the data. For example, to test whether there is any difference between the two conditions we can use the *Wilcoxon test*. To do this, we take each condition and calculate the sum of ranks, and subtract the least value it could have (that is, $1+2+3+4+5=15$ for condition A, $1+2+3+4+5+6=21$ for condition B), giving the statistic *U*:

	rank sum		least		*U*
condition A:	$(2+3+1+7+4)$	$-$	15	$=$	2
condition B:	$(8+6+10+11+9+5)$	$-$	21	$=$	28

In fact, the sum of these two *U* statistics, $2 + 28 = 30$, is the product of the number of data values in each condition 5×6. This will always happen and so one can always get away with calculating only one of the *U*. Finally, we then take the smaller of two *U* values and compare it with a set of *critical values* in a book of statistical tables, to see if it is unusually small. The table is laid out dependent on the number of data values in each condition (five and six). The critical value at the 5% level turns out to be 3. As the smallest statistic is smaller than this, we can *reject the null hypothesis* and conclude that there is likely to be a difference between the conditions. To be precise, it says that there is only a 1 in 20 (5%) chance that the data happened by chance. In fact the test is right – the authors constructed random data in the range 1–100 and then subtracted 10 from each of the values in condition A.

An example: evaluating icon design

- Hypothesis : User will remember the natural icons more easily than the abstract ones.
- Null hypothesis is no difference between recall of the icon types.

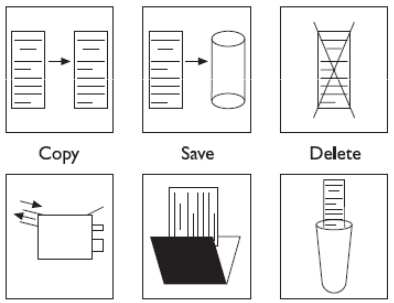

Figure 9.3 Abstract and concrete icons for file operations

An example: evaluating icon design

- ตัวแปรต้น : varying the style of icon (natural and abstract)
- ตัวแปรตาม : the number of mistakes in selection and the time taken to select an icon.
- ตัวแปรควบคุม : อุปกรณ์ที่ใช้ทดลอง (computer, mouse, keyboard), operating system, light, size of icon, color tone of icon, etc.
- A between-subjects experiment would remove any learning effect for individual participants, but it would be more difficult to control for variation in learning style between participants.
- => A within-subjects design is preferred, with order of presentation controlled.
- The user is presented with a task (say 'delete a document') and is required to select the appropriate icon.

An example: evaluating icon design

- To *avoid learning effects* from icon position, the placing of icons in the block can be randomly varied on each presentation. Users are divided into two groups with each group taking a different starting condition.
- Measure the time taken to complete the task and the number of error made.

Table 9.2 Example experimental results

Participant number	Presentation order	(N (s
1	AN	
2	AN	
3	AN	
4	AN	
5	AN	
6	NA	
7	NA	
8	NA	
9	NA	
10	NA	
mean (μ)		
s.d. (σ)	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$	
Student's *t*		

< 5 mins

≈ 20 mins

Mean diff = 52 s

s.e.d. 117

0.32 (n.s.)

s.e. 4.55 =s.d./√10

5.78 (p<1%, two tailed)

Table 4
Critical Values of *t* for Two-Tailed *t* Test[1]

df	.05 level	.01 level	.001 level
1	12.706	63.657	636.619
2	4.303	9.925	31.598
3	3.182	5.841	12.941
4	2.776	4.604	8.610
5	2.571	4.032	6.859
6	2.447	3.707	5.959
7	2.365	3.499	5.405
8	2.306	3.355	5.041
9	2.262	3.250	4.781
10	2.228	3.169	4.587
11	2.201	3.106	4.437
12	2.179	3.055	4.318
24	2.064	2.797	3.745
25	2.060	2.787	3.725
26	2.056	2.779	3.707
27	2.052	2.771	3.690
28	2.048	2.763	3.674
29	2.045	2.756	3.659
30	2.042	2.750	3.646
40	2.021	2.704	3.551
60	2.000	2.660	3.460
120	1.980	2.617	3.373
infinity	1.960	2.576	3.291

And we used a within-subject design, there is another independent variable – the participant. => we should use analysis of variance (ANOVA)

- Empirical methods: experimental evaluation
  - Studies of groups of users
    - The participant groups ; Ex. 3 exp x 10 participants, take time longer than single-user.
    - The experimental task ; the task also depends on the nature of the groupware system.
    - Data gathering
    - Analysis
    - Field study with groups

- Observational techniques
  - Think aloud and cooperative evaluation
    - Observation where the user is asked to talk through what he is doing as he is being observed. Ex. Describing what he believes is happening, why he takes an action, what he is trying to do.
    - The evaluator can clarify point of confusion.
  - Protocol analysis
    - Paper and pencil
    - Audio recording
    - Video recording
    - Computer logging (Ex. Record user actions at a keystroke)
    - User notebook (Ex. Participants be asked to keep logs of activity/problems.)

- Observational techniques
  - Automatic protocol analysis tools

Figure 9.4 EVA: an automatic protocol analysis tool. Source: Wendy Mackay

  - Post-task walkthroughs
    - To reflect the participant's actions back to them after the event, by asking comment, or directly question.
    - This is only way to obtain a subjective viewpoint on the user's behavior.

- Query techniques
  - Interview
    - The level of questioning can be varied to suit the context and that the evaluator can probe the user more deeply on interesting issues as they arise.
  - Questionnaires

    Less flexible than interview technique, questions are fixed in advance. But it can reach wider participant group and take less time.
    - General (Ex. Age, sex, occupation, experience, etc.)
    - Open-ended (Ex. Can you suggest any improvements to the interface?)
    - Scalar (judge a specific statement on a numeric scale)

      It is easy to recover from mistakes.
      Disagree 1 2 3 4 5 Agree

- Query techniques
  - Questionnaires
    - Multi-choice

      How do you most often get help with the system (tick one)?
      Online manual ❑
      Contextual help system ❑

      Please rank the usefulness of these methods of issuing a command (1 most useful, 2 next, 0 if not used).
      Menu selection ❑
      Command line ❑
      Control key accelerator ❑
      Expert system ❑
      Online help system ❑
      Compiler ❑

- Evaluation through monitoring physiological responses
  - Eye tracking for usability evaluation
    - Number of fixations
    - Fixation duration
    - Scan path

Figure 9.5 Eye-tracking equipment. Source: Courtesy of J. A. Renshaw

- Evaluation through monitoring physiological responses
  - Physiological measurements
    - Heart activity; blood pressure, volume and pulse.
    - Activity of the sweat glands; galvanic skin response(GSR)
    - Electrical activity in muscle; electromyogram (EMG)
    - Electrical activity in the brain; electroencephalogram(EEG)

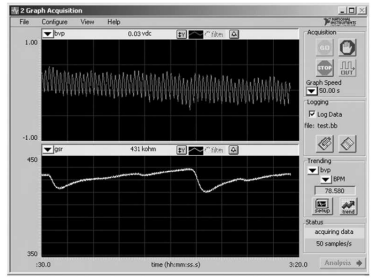

Figure 9.8 Output of monitoring pulse rate (above) and skin conductivity (below). Source: Screen shot courtesy of Dr R. D. Ward; frame source: National Instruments BioBench software

Choosing an evaluation method

- Factors distinguishing evaluation techniques
  - Design vs. implementation
    - The main distinction is in evaluation in implementation is a physical artifact exists, something concrete can be tested.
  - Laboratory vs. field studies
    - Field studies retain naturalness of the user's environment but do not allow control over user activity.
    - Ideally the design process should include both styles of the evaluation.

- Factors distinguishing evaluation techniques
  - Subjective vs. objective
    - Bias in subjective techniques should be recognized and avoided by using more than one evaluator.
    - Objective techniques should produce repeatable results, which are not dependent on the persuasion of the particular evaluator.
  - Qualitative vs. quantitative measures
    - Quantitative measurement is usually numeric and can be easily analyzed using statistical techniques.
    - Qualitative measurement is opposite, but can provide important detail that cannot be determined from numbers.

- Factors distinguishing evaluation techniques
  - Information provided
    - Low-level information (Ex. Which font is most readable) => an experiment can be designed to measure a particular aspect of the interface.
    - Higher-level information (Ex. Is the system usable?) => can be gathered using questionnaire and interview techniques, which provide a more general impression of the user's view of the system.
  - Immediacy of response
    - Ex. Think aloud, record user's behavior at the time of the interaction itself, post-task walkthrough.

- Factors distinguishing evaluation techniques
  - Intrusiveness
    - Most immediate evaluation techniques run the risk of influencing the way the user behaves.
  - Resource
    - Resources to consider include equipment, time money, participants, expertise of evaluator and context.
    - Some decisions are forced by resource limitations. Ex. It is not possible to produce a video protocol without access to a video camera.

- A classification of evaluation techniques

Table 9.4 Classification of analytic evaluation techniques

	Cognitive walkthrough	Heuristic evaluation	Review based	Model based
Stage	Throughout	Throughout	Design	Design
Style	Laboratory	Laboratory	Laboratory	Laboratory
Objective?	No	No	As source	No
Measure	Qualitative	Qualitative	As source	Qualitative
Information	Low level	High level	As source	Low level
Immediacy	N/A	N/A	As source	N/A
Intrusive?	No	No	No	No
Time	Medium	Low	Low–medium	Medium
Equipment	Low	Low	Low	Low
Expertise	High	Medium	Low	High

- A classification of evaluation techniques

Table 9.5 Classification of experimental and query evaluation techniques

	Experiment	Interviews	Questionnaire
Stage	Throughout	Throughout	Throughout
Style	Laboratory	Lab/field	Lab/field
Objective?	Yes	No	No
Measure	Quantitative	Qualitative/ quantitative	Qualitative/ quantitative
Information	Low/high level	High level	High level
Immediacy	Yes	No	No
Intrusive?	Yes	No	No
Time	High	Low	Low
Equipment	Medium	Low	Low
Expertise	Medium	Low	Low

- A classification of evaluation techniques

Table 9.6 Classification of observational evaluation techniques

	Think aloud[1]	Protocol analysis[2]	Post-task walkthrough
Stage	Implementation	Implementation	Implementation
Style	Lab/field	Lab/field	Lab/field
Objective?	No	No	No
Measure	Qualitative	Qualitative	Qualitative
Information	High/low level	High/low level	High/low level
Immediacy	Yes	Yes	No
Intrusive?	Yes	Yes[3]	No
Time	High	High	Medium
Equipment	Low	High	Low
Expertise	Medium	High	Medium

1 Assuming a simple paper and pencil record
2 Including video, audio and system recording
3 Except system logs

- A classification of evaluation techniques

Table 9.7 Classification of monitoring evaluation techniques

	Eye tracking	Physiological measurement
Stage	Implementation	Implementation
Style	Lab	Lab
Objective?	Yes	Yes
Measure	Quantitative	Quantitative
Information	Low level	Low level
Immediacy	Yes	Yes
Intrusive?	No[1]	Yes
Time	Medium/high	Medium/high
Equipment	High	High
Expertise	High	High

1 If the equipment is not head mounted

Consent form

- Depending on the agency or institution overseeing the research, participants are usually required to sign a consent form prior to testing.
- The goal is to ensure participants know
  - that their participation is voluntary,
  - that they will incur no physical or psychological harm,
  - that they can withdraw at any time, and
  - that their privacy, anonymity, and confidentiality will be protected.

Consent form

Project Title: Real Phone -
Evaluation of its Usability
Investigator(s): Saul Greenberg

Sponsoring Company: Saul Greenberg Consulting, 220 6087

This consent form should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, feel free to ask. Please take the time to read this carefully and to understand any accompanying information.

Description and conditions of the study.
IBM has developed a software product called "RealPhone" that is a software equivalent of a telephone. The purpose of this study is to discover what problems people may have when using it. You have been chosen to help us both because you have used Microsoft Windows, but you have not used the RealPhone product before.

You will be given the RealPhone software, and asked to perform a series of tasks with it. The entire process will take about one hour. As you perform the tasks, you will be asked to 'think aloud', where you say what you are thinking as you are doing it. You will also be asked to fill out a short series of questions before, during, and after the study. At the end, we will interview you about your experiences. During this, we will videotape what you are doing on the screen and capture your voice (although your face will not be captured on tape).

You may feel uncomfortable or awkward if you have trouble doing some of these tasks. However, we are testing the system, not you. If you have any problems with any of the tasks we ask you to do, that is exactly what we are looking for because we can then repair these problem areas and improve the product. Still, if you find this test objectionable in any way, you are free to quit at any time.

When discussing the results of this study, we will never refer to you by name or show your picture. In fact, we do not record your name on any of our records except on this sheet. Records and videotapes will be kept in a secure area and will only be disclosed to other members of the evaluation and development team. No other people will have access to this material.

Your participation is completely voluntary, without any payment.

Your signature on this form indicates that you have understood to your satisfaction your participation in the research project and that you agree to participate as a subject. In no way does this waive your legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. You are free to withdraw from the study at any time. Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please contact: Saul Greenberg, 220 6087

Participant's Signature
Investigator Signature
Witness' Signature
A copy of this consent form has been given to you to keep for your records and reference.

What this is
Study purpose
What will happen
Risks
Confidentiality
What they are signing

Saul Greenberg

- The experiment begins.
- The experimenter greets each participant, introduces the experiment, and usually asks the participants to sign consent forms.
- Often, a brief questionnaire is administered to gather demographic data and information on the participants' related experience.
- This should take just a few minutes.
- The apparatus is revealed, the task explained and demonstrated.
- Practice trials are allowed, as appropriate.

Summary

- Aim of evaluation is to test the functionality and usability of the design and to identify and rectify any problems.
- A design can be evaluated before any implementation work has started, to minimize the cost of early design errors.
- Query techniques provide subjective information from the user. For objective information, physiological monitoring can capture the user's physical responses to the system.
- The choice of evaluation method is largely dependent on what is required of the evaluation.