Chapter 8: User Interface Evaluation

The last remaining part in the cycle of UI (interactive software) development is the "Evaluation" stage. Even if the developers may have strived to adhere to various HCI principles, guidelines and rules, and applied the latest toolkits and implementation methodologies, the resulting UI or software is most probably not problem-free. Frequently, careful considerations in interaction and interface design may not even have been carried out in the first place. Aside from the fact that there may be things that the developer oversaw to consider, the overall development process was to be a gradual refinement process to begin with, where the next refinement stages would be based on the evaluation results of the previous rounds. In this chapter, we will present several methods and examples of evaluation for user interfaces.

8.1 Evaluation Criteria

When evaluating the interaction model and interface, there are largely two criteria. One is the **usability** and the other **user experience (UX)**. Simply put, usability refers to the ease of use and learnability of the user interface (we come back to UX later in the section) [12]. Usability can be measured in two ways, **quantitatively** or **qualitatively**.

Quantitative assessment often involves task performance measurements. That is, we assume that an interface is "easy to use and learn" (good usability) if the subject (or a reasonable pool of subjects) is able to show some (absolute) minimum user performance on typical application tasks. The assessment of a given new interface is better made in a comparative fashion against some

nominal or conventional interface (in terms of relative performance edge). Popular choices of such performance measure are task completion time, task completion amount in a unit time (e.g. score), and task error rate. For example, suppose we would like to test a new motion based interface for a smart phone game. We could have a pool of subjects play the game, using both the conventional touch based interface and also the newly proposed motion based one. We could compare the score and assess the comparative effectiveness of the new interface. The underlying assumption is that task performance is closely correlated to the usability (ease of use and learnability). However, such an assumption is quite arguable. In other words, task performance measures, while quantitative, only reveals the aspect of efficiency, or merely the aspect of ease of use, not necessarily the entire usability. The aspect of learnability should be and can be assessed in a more explicit way, by measuring the time and effort (e.g. memory) for users to learn the interface. The problem is that it is difficult to gather a "homogeneous" pool of subject with similar backgrounds (in order to make the evaluation fair). Measuring the learnability generally is likely to introduce much more biasing factors such as differences due to educational/experiential/cultural background, age, gender, etc. Finally, quantitative measurements in practice cannot be applied to all the possible tasks for a given application and interface. Usually a very few representative tasks are chosen for evaluation. This sometimes makes the evaluation only partial.

To complement the shortcomings of the quantitative evaluation, **qualitative evaluations** often are conducted together with the quantitative analysis. In most cases, quantitative evaluations

amount to conducting a "usability" survey, asking usability related questions to a pool of subjects after having them experience the interface. A usability survey often includes questions involving the ease of use, ease of learning, fatigue, simple preference, and other questions specific to the given interface. NASA TLX (Task Load Index, Figure 8.1) and the IBM Usability Questionnaire (Figure 8.2) are examples of the often used semi-standard questionnaires for this purpose [3][4][10].

[Figure 8.1] Excerpts from the NASA TLX Usability Questionnaire [3][10]. NASA Task Load Index mehtod assess world load on 7 point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scale.

[Figure 8.2] Excerpts from the IBM Usability Questionnaire for computer systems [7].

User experience (UX) is the other important aspect of interface evaluation. There is no precise definition for UX. It is generally accepted that the notion of user experience is "total" in the sense that it is not just about the interface but also something about the whole product/application and even extend to the product family (such as the Apple products or MS Office). It is also deeply related to the user's emotion and perception that result from the use or

anticipated use of the application (through the given interface) [2]. Such an affective response is very much dependent on the context of use. Thus UX evaluation involves a more comprehensive assessment on the emotional response, under variety of usage contexts and across a family of products/applications/interfaces (see Figure 8.3). A distinction can be made between usability methods that have the objective of improving human performance, and user experience methods that have the objective of improving user satisfaction with achieving both the pragmatic and hedonic goals [1]. Note that the notion of UX includes usability: that is, "usually" high UX translates to high usability and high emotional attachment.

[Figure 8.3] Various aspects to be considered in totality for assessing user experience (UX).

8.2 Evaluation Methods

Whether it is for the user experience or more narrow usability or whether for the qualitative feelings or quantitative performance, there exists a variety of evaluation methods. A given method may be general and applicable to many different situations and objectives, or more specific and fitting for a particular criterion or usage situation. Overall, an evaluation method can be characterized by the following factors:

- Timing of analysis (e.g. throughout the application development stage: early, middle,

late/after),

- Type and number of evaluators (e.g. several HCI experts vs. 100's of domain users),
- Formality (e.g. controlled experiment or quick and informal assessment), and
- Place of the evaluation (laboratory vs. in-situ field testing).

8.2.1 Focus Interview / Enactment / Observation Study

One of the easiest and straightforward evaluation methods is to simply **interview** the actual/potential users and **observe their interaction behavior** either with the finished product or through a simulated run. The interview can be conducted in a simple question-&-answer form, and can involve an actual usage of the given system/interface. Depending on the stage of the development at which the evaluation takes place, the application or interface may not be ready for such a full-fledged test drive. Thus, a simple paper/digital mock-up may be used so that a particular usage scenario is enacted for which the interview can be based on. While mock-ups provide a tangible product and thus an improved feel for the system/interface (vs. a mere rough paper sketch), at an early stage of the development, important interactive features may not have been implemented as yet. In this case, a **"Wizard-of-Oz"** type of testing is often employed, where a human administrator fakes the system response "behind the curtain." User interaction behaviors during the test trials or simulation runs are recorded or video-taped for more detailed post-analysis.

[Figure 8.4] Interviewing a subject upon simulating the usage of the interface with a mock-up.

The interview is often **"focused"** on particular user groups (e.g. elderly) or features of the system/interface (e.g. information layout) to save time. One particular interviewing technique is called the **"cognitive walkthrough"** in which the subject (or expert) is asked to "speak aloud" his thought process. In this case, the technique is focused on investigating for any gap between the interaction model of the system and that of user. We can deduce that cognitive walkthroughs are fit for evaluation at a relatively earlier stage of design, namely interaction modeling or interface selection (vs. specific interface design). Another notable variation of the actual usage based testing is the **"Can you break this?"** type of testing in which the subject is given the mission to explicitly expose interface problems e.g. by demonstrating interface flaw and interface design related "bugs."

[Figure 8.5] A cognitive walkthrough with the interviewer.

Note that the interview/simulation method, due to its simplicity, can be used not only for

evaluation but also for interaction modeling and exploration of alternatives at the "early" design stage. We have already seen in Chapter 4 design tools such as storyboards, wire-framing and GOMS which can be used in conjunction with users or experts for simultaneous analysis and design. The user interviewing/observation technique, being somewhat free form, is easy to administer, but not structured to be comprehensive. The following table summarizes the characteristics of the interview/simulation/observation approach.

[Table 8.1] Summary: Interview, usage and observation method

Evaluators / Size	Actual users / Medium sized (10~15)		
Type of evaluators	Focused (e.g. by expertise, age group, gender, etc.)		
Formality	Usually informal (not controlled experiment)		
Timing and Objectives	Stage	Objective	Enactment Method
	Early	Interaction model and flow	Mock-up / Wizard of Oz
	Middle	Interface selection	Mock-up / Wizard of Oz Partial simulation
	Late/After	Interface design issues (look and feel such as aesthetics, color, contrast, font size, icon location,	Simulation Actual system

		labeling, layout, etc.)	
Easy to administer / Free Form, but not structured nor comprehensive			

8.2.2 Expert Heuristic Evaluation

Expert heuristic evaluation is very similar to the interview method. The difference is that the evaluators are HCI experts and the analysis is carried out against a pre-prepared HCI guideline, hence called heuristics. For instance, the guideline can be general or more specific (Chapter 2), with respect to application genre (e.g. for games), cognitive/ergonomic load, corporate UI design style (e.g. Android UI guideline), and etc. The directions or particular themes of the heuristics are chosen by the underwriter. The following lists Nielsen's 10 general UI heuristics. Note that these guidelines are almost same to the general principles/guidelines introduced in Chapter 1~2 and used for interaction/interface design.

- **Visibility of system status:** The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
- **Match between system and the real world:** The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

- **User control and freedom:** Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
- **Consistency and standards:** Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
- **Error prevention:** Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.
- **Recognition rather than recall:** Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
- **Flexibility and efficiency of use:** Accelerators—unseen by the novice user—may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.
- **Aesthetic and minimalist design:** Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

- **Help users recognize, diagnose, and recover from errors:** Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
- **Help and documentation:** Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

In the far left and middle columns of Table 8.2, we show evaluation heuristics specifically derived for evaluating the initial design of "No Sheets" (done in Chapter 4). The heuristics were derived by the developer who identified, among very many, the more important principles and guidelines to follow for this particular application. The right column shows partial results of applying these evaluation heuristics. This way, the evaluation was carried out by third party HCI expert efficiently by paying a particular attention to those heuristics.

[Table 8.2] Evaluation heuristics derived specifically for evaluating "No Sheets" and its application results.

Heuristic	Specifics (examples)	Evaluation results (partial)
System status	Does the user understand what is going as the song is played?	While playing, the tempo and whether it is being played, fast

	(e.g. part of the song is being played, current operation, etc.)	forwarded, or reviewed, is not clearly shown.
Display layout	Is the information laid out and positioned properly (e.g. chords, beat, lyrics), Is the color coding and icon design proper for fast recognition?	The colors are too raw (tiring to the eyes). Landscape mode is preferred (vs. portrait). The icon designs for "fast forward", "review" are not familiar.
Interaction / Contents model	Are there all the essential functionalities for this application? Are necessary functions or accessible or information displayed at different interaction points?	Tempo control, fast forward, and review are not possible during play. Information per measures are needed.
Ergonomic consideration / User characteristics / Operating environment	Assess readability, color contrast, GUI object size. Also assess if easily operable in a typical/various usage situation (for piano, guitar, etc.)	A better color contrast is needed between different types of information. Provision is needed for long lyrics. Landscape mode is more desirable.
Input/Output method	Assess interface methods:	Beat sound is too high pitched.

	conveying the beat (beat number, sound), setting the tempo, selecting the song, etc.	Suggest dragging for fast forward and review functions.
Consistency / Standards	Evaluate consistency with actual sheet music and Android design guideline.	A more common choice or design of icons is needed.
Prevention of errors	Is the interaction modeled designed such that to minimize error? Is it possible to easily undo?	Explicitly deactivate the play button when there is no song selected.
Aesthetics	Evaluate simplicity and overall attractiveness.	Mostly simple except for using too much primary colors
Help	Is there sufficient help and guides for the beginner?	Need more detailed guide and introduction.

[Figure 8.6] The initial (left) and redesigned (right) "play" activity/layer for No Sheets: the new design after evaluation uses a landscape mode is used and less primary colors. The icons for fast forward and review are changed to more conventional ones and the current tempo is shown on top.

The expert heuristic evaluation is one of the most popular methods of UI evaluation because it is quick and dirty and relatively cost effective. Only a few (typically 3~5) UI and domain experts are typically brought in to evaluate the UI implementation in the late stage of the development or even against a finished product. The disadvantage of the expert review is that the feedback from the user is absent as the HCI expert may not understand the needs of the actual users. On the other hand, the small sized evaluator pool is compensated by their expertise.

[Table 8.3] Summary of the expert review method.

Evaluators / Size	HCI Experts / Small sized (3~5)		
Type of evaluators	Focused (Experts on Application specific HCI rules, Corporate specific design style, User ergonomics, etc.), Interface consistency		
Formality	Usually informal (not controlled experiment)		
Timing and Objectives	Stage	Objective	Enactment Method
	Middle	Interface selection	Scenarios

			Storyboards Interaction Model
	Late/After	Interface design issues (look and feel such as aesthetics, color, contrast, font size, icon location, labeling, layout, etc.)	Simulation Actual system
Easy and quick, but prior heuristics assumed to exist and no actual user feedback reflected			

8.2.3 Measurement

In contrast to interviews and observation, measurement methods attempt to indirectly quantify the goodness of the interaction/interface design with **a score** through representative task performance (quantitative) or quantified answers from carefully prepared subjective surveys (qualitative).

Typical indicators for quantitative task performance are the task completion time, score (or amount of task performance in unit time) and errors (produced in unit time). For example, for a mobile game, a representative task might be to "invoke the given game, log in, and reach the main screen." Another example task, for "No Sheets," would be to "invoke the application, load the music file, and set the tempo." Task performance measurement is only meaningful when compared to the nominal/reference case. Thus, two measurements must be made between the

nominal and "new" design, and statistical analysis is applied to derive any meaningful and significant differences between the two measurements. It is generally accepted that it is much more feasible to gather a sufficiently homogenous yet relatively smaller subject pool for "physical/cognitive" task performance measurement without much bias or variation.

[Figure 8.7] A case of a task performance measurement: (1) nominal: a game interface using a keyboard, and (2) new: a game interface using a new controller. Task completion time is measured for navigating a maze using the respective interface and compared to indirectly assess the ease of interaction.

On the other hand, numerical scores can be obtained from surveys. Surveys are used because many aspects of usability or user experience are based on user perception which is not directly measurable. However, answers to user perception qualities are highly variable and much more susceptible to user's intrinsic backgrounds. To reduce such biases, a few provisions can be made, for example using a large number of subjects (e.g. more than 30 people), using an odd-leveled (5 or 7) answer scale (also known as the Likert scale [Likert] so that there always exist the middle level answer, carefully wording and explaining the survey question for clarity and understanding (more guidelines in Table 8.4). Even though the result of the survey is a numerical score, the nature of the measurement is still qualitative because survey questions usually deal with user

perception qualities. Similarly to the task performance case, comparative survey, against the nominal case, is recommended.

[Table 8.4] Guidelines for a good survey.

Minimize the number of questions	Too many questions results in fatigue and hence unreliable responses.
Use an odd level scale, 5 or 7 (or Likert Scale)	Research has shown odd answer levels with mid value with 5 or 7 levels produces the best results.
Use consistent polarity	E.g. negative responses correspond to level 1 and positive to 7 and consistently so throughout the survey.
Make questions compact and understandable	Questions should be clear and easy to understand. If difficult to convey the meaning of the question in compact form, the administrator should verbally explain.
Give subjects compensation	Without compensation, subjects will not do one's best or perform the given task reliably.
Categorize the questions	For easier understanding and good flow, questions of the same nature should be grouped and answered in block, e.g. answer "ease of use" related questions,

	then "ease of learning" and so on.

Both types of measurement experiments can optionally be run over a long period of time, especially when memory performance and familiarity aspect is involved. For instance, to assess ease of learning of an interface, the task performance can be measured over weeks to see how quickly the user recalls how to operate the interface and produce higher performance.

Another variation is with the place of the evaluation. When testing with the finished product, it is best to conduct the usage test at the actual place of usage, outside the laboratory (e.g. at the office, at home, on the street, etc.). However, as expected, it is often very difficult operational-wise to conduct the measurement or testing at the actual place of interaction. Even if it was possible there are many uncontrollable factors that might affect the outcome of the testing (e.g. having to test in front of other people). To isolate and prevent these possible biases, the testings are often conducted in a laboratory setting as well with carefully selected homogenous pool of subjects.

With the advent of the smart phones and their ubiquity, the in-situ field testing is gaining great popularity [11]. Applications can collect user interaction information in the background upon particular interaction events and be analyzed in a batch process. While the same danger exists with respect to the environmental biases, they are often mitigated by the high number of subjects (e.g. users of smart phones and apps). Some research have shown that there is very little

difference in the analysis/evaluation results between the controlled laboratory studies and in-situ field studies [5]. However depending on the applications (especially those for which typical usage situations cannot easily be recreated in the laboratory) [6].

In fact, in addition to the need to carefully construct the survey, measurement experiments require a meticulous operational logistics to be as fair and bias free as possible, starting from recruitment and screening of the subjects, pre-training of them, compensation and obtaining consent, choosing the right independent and dependent variables and applying the right statistical analysis methods to the resulting data. The detail of such "design of experiments" is beyond the scope of this book and we refer you to the related literatures [DEX]. Despite the higher reliability of the evaluation results, significant amount of efforts are needed to prepare and administer the measurement type of interface evaluation

[Table 8.5] Summary of the measurement method.

Evaluators / Size	Potential/typical users / Medium-Large sized (10~more than 50)		
Type of evaluators	Balanced and homogeneous pool of subjects (users of the system) (Gender, age, educational background, relevant skills, etc.)		
Formality	Can be formal controlled experiment or informal		
Place	Laboratory or in-situ field		
Timing and Objectives	Stage	Objective	Enactment Method

	Late/After	Interface design issues (look and feel such as aesthetics, color, contrast, font size, icon location, labeling, layout, etc.)	Simulation Actual system
More reliable result but generally time consuming to prepare and conduct the process			

8.2.4 Safety and Ethics in Evaluation

Most HCI evaluation involves simple interviews and or carrying out "simple" tasks using of paper mock-ups, simulation systems or prototypes. Thus, safety problems rarely occur. However, precautions are still needed. For example, even interviews can become long and time consuming and causing much fatigue on to the subject. Certain seemingly harmless tasks may bring about unexpected harmful effects both physically and mentally. Therefore, evaluations must be conducted on volunteers and with signed consents. Even with signed consents, the subjects take the right to discontinue the evaluation task any time. The purpose and the procedure should be sufficiently explained and made understood to the subjects prior to any experiments. Many organizations run what is called the Institutional Review Board (IRB) will review the one's evaluative experiments to ascertain safety and rights of the subjects. It is best to consult or obtain permission from the IRB for when there is even a small doubt of some kind of effect to the subjects during the experiments.

8.3 Summary

We have looked at various methods for evaluating the interface at different stages in the development process. As already emphasized, even though all the provisions and knowledge may have been put to use to create the initial versions of the UI, many compromises may be made during the actual implementation, resulting in a product somewhat different from what was originally intended at the design stage. It is also quite possible that during the course of the development, the requirements simply change. This is why the explicit evaluation step is a must and in fact the whole design-implement-evaluate cycle must ideally be repeated at least few times until a stable result is obtained.

References

[1] Bevan, N. (2008) UX, Usability and ISO Standards. Values, Value and Worth workshop, CHI 2008.

[2] DIS, ISO. "9241-210: 2010. Ergonomics of human system interaction-Part 210: Human-centered design for interactive systems." International Organization for Standardization (ISO), (2009).

[3] Hart, Sandra G., Land, Steve, and Lowell E. "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research." *Human mental workload* 1.3 (1988): 139-183.

[4] Jang, Bong-gyu, and Kim, Gerard J. "Evaluation of grounded isometric interface for whole-body navigation in virtual environments." *Computer Animation and Virtual Worlds* (2013).

[5] Kaikkonen, A., Kallio, T., Kekalainen, A., Kankainen, A. and Cankar, M. (2005) Usability testing of mobile applications: A comparison between laboratory and field testing. Journal of Usability Studies, 1(1):4-16.

[6] Kjeldskov, J., Skov, M.B., Als, B.S. and Høegh, R.T. (2004) Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. In Proceedings of the 6th International Mobile HCI 2004 conference. LNCS, Springer-Verlag.

[7] Lewis, James R. "IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use." *International Journal of Human-Computer Interaction* 7.1 (1995): 57-78.

[8] Likert, Rensis. "A Technique for the Measurement of Attitudes." *Archives of Psychology* 22.140

(1932): 1–55.

[9] Nielsen, Jakob. "Enhancing the explanatory power of usability heuristics." *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, (1994): 152-158.

[10] NASA, "NASA Task Load Index" http://humansystems.arc.nasa.gov/groups/tlx/downloads/TLXScale.pdf, (2013).

[11] Rowley, D. E. (1994) Usability Testing in the Field: Bringing the Laboratory to the User. In Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press.

[12] Wikipedia, "Usability" http://en.wikipedia.org/wiki/Usability, (2013).

[13] Wikipedia, "Wizard of Oz experiment" http://en.wikipedia.org/wiki/Wizard_of_Oz_experiment, (2013).