Human Computer Interaction

Fundamentals and Practice [SWE - 431]

Gerard Jounghyun Kim

Chapter: 8
User Interface Evaluation

Mahfuzur Rahman Emon
Lecturer, IICT, SUST

Evaluation Criteria

When evaluating the interaction model and interface, there are largely two criteria

- Usability: Refers to the ease of use and learnability of the user interface
- User Experience

Usability: Refers to the ease of use and learnability of the user interface. Can be measure in two ways

- Quantitatively
- Qualitatively

Quantitative Assessment:

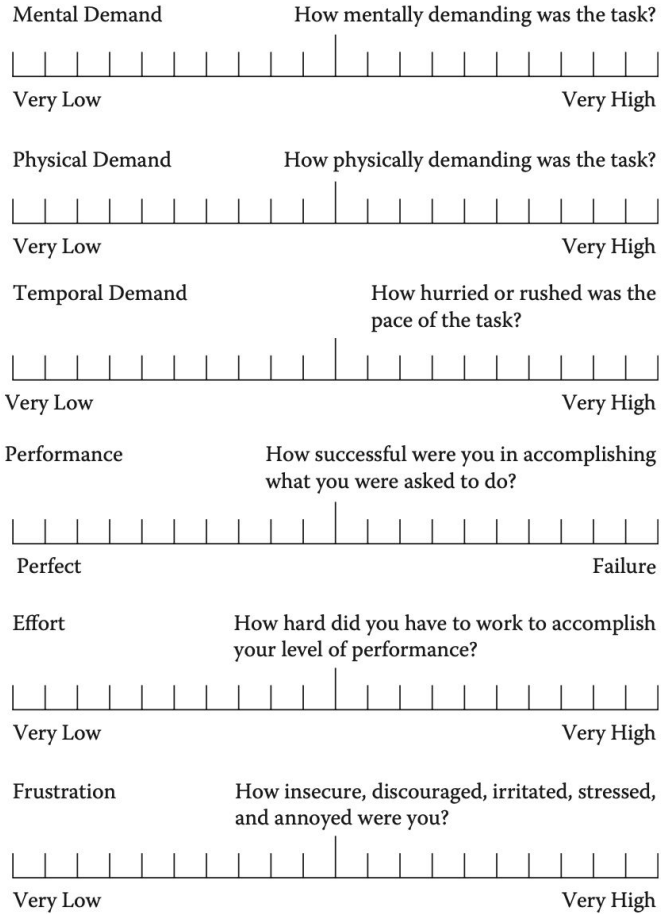
- Involves task-performance measurements. An interface is “easy to use and learn” if the subject/subjects is/are able to show some minimum user performance on typical application tasks.
- Popular choices of such performance measures are task completion time, task completion amount in a unit time (e.g., score), and task error rate
- For example, we would like to test a new motion-based interface for a smartphone 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), such an assumption is quite arguable.
- Task-performance measures, while quantitative, only reveal the aspect of efficiency (or merely the aspect of ease of use) and not necessarily the entire usability.

Quantitative Assessment:

- 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 subjects with similar backgrounds (in order to make the evaluation fair)
- Measuring the learnability is generally 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.

Qualitative Evaluations:

- To complement the shortcomings of the quantitative evaluation, qualitative evaluations often are conducted together with the quantitative analysis.
- Conducting a usability survey, posing usability-related questions to a pool of subjects after having them experience the interface.
- Includes questions involving the ease of use, ease of learning, fatigue, simple preference, and other questions specific to the given interface.
- Ex. NASA TLX (Task Load Index) and the IBM Usability Questionnaire



NASA TLX Usability Questionnaire

The NASA Task Load Index method assesses the workload on a seven-point scale. Increments of high, medium, and low estimates for each point result in 21 gradations on the scale

1. Overall, I am satisfied with how easy it is to use this system.

Strongly Agree **Strongly Disagree**

Comments: 1 2 3 4 5 6 7

2. It was simple to use this system.

Strongly Agree 1 2 3 4 5 6 7 **Strongly Disagree**

Comments:

3. I could effectively complete the tasks and scenarios using the system.

Strongly Agree 1 2 3 4 5 6 7 **Strongly Disagree**

Comments:

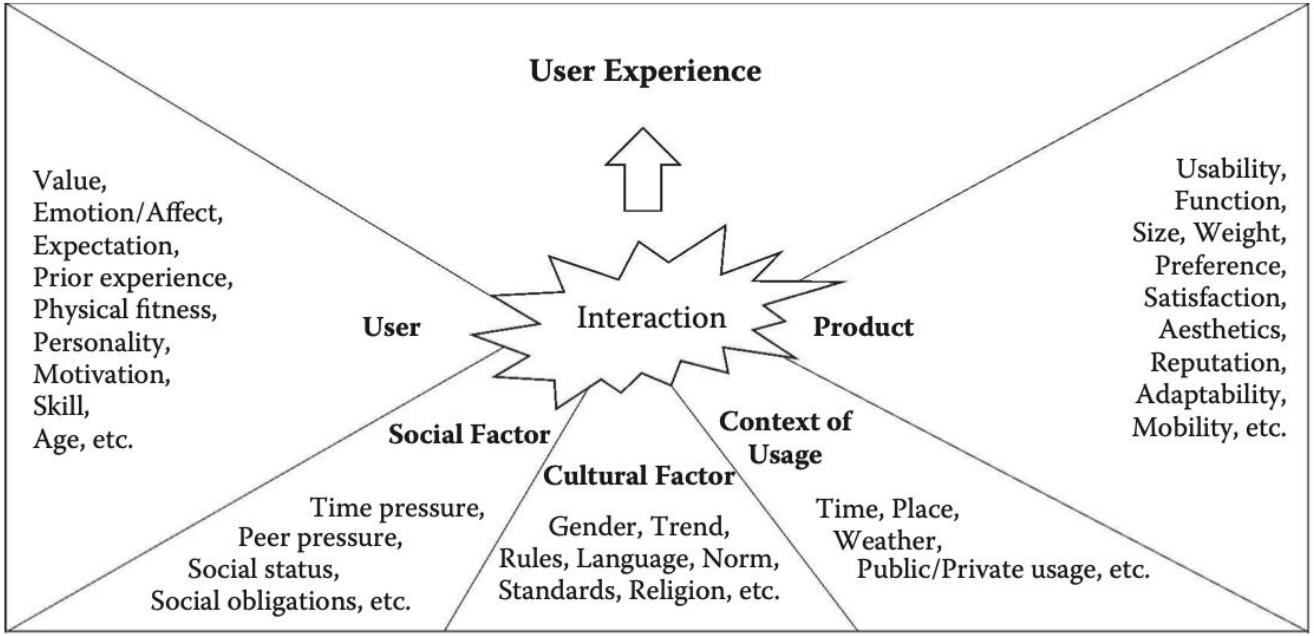
4. I was able to complete the tasks and scenarios quickly using this system.

Strongly Agree 1 2 3 4 5 6 7 **Strongly Disagree**

Comments:

IBM Usability Questionnaire for computer systems.

User Experience: There's no precise definition, A holistic concept encompassing more than just the interface, extending to the entire product or application, and even the product family. It involves the user's emotions and perceptions arising from using the application through a given interface.



Various aspects to be considered in totality for assessing user experience (UX).

Evaluation Methods

Whether it is for the *user experience* or more narrow *usability*, or whether for the *qualitative feelings* or *quantitative performance*, there is a variety of evaluation methods.

A given method may be general and applicable to many different situations and objectives, or it may be more specific and fitting for a particular criterion or usage situation. Overall, an evaluation method can be characterized by the following factors:

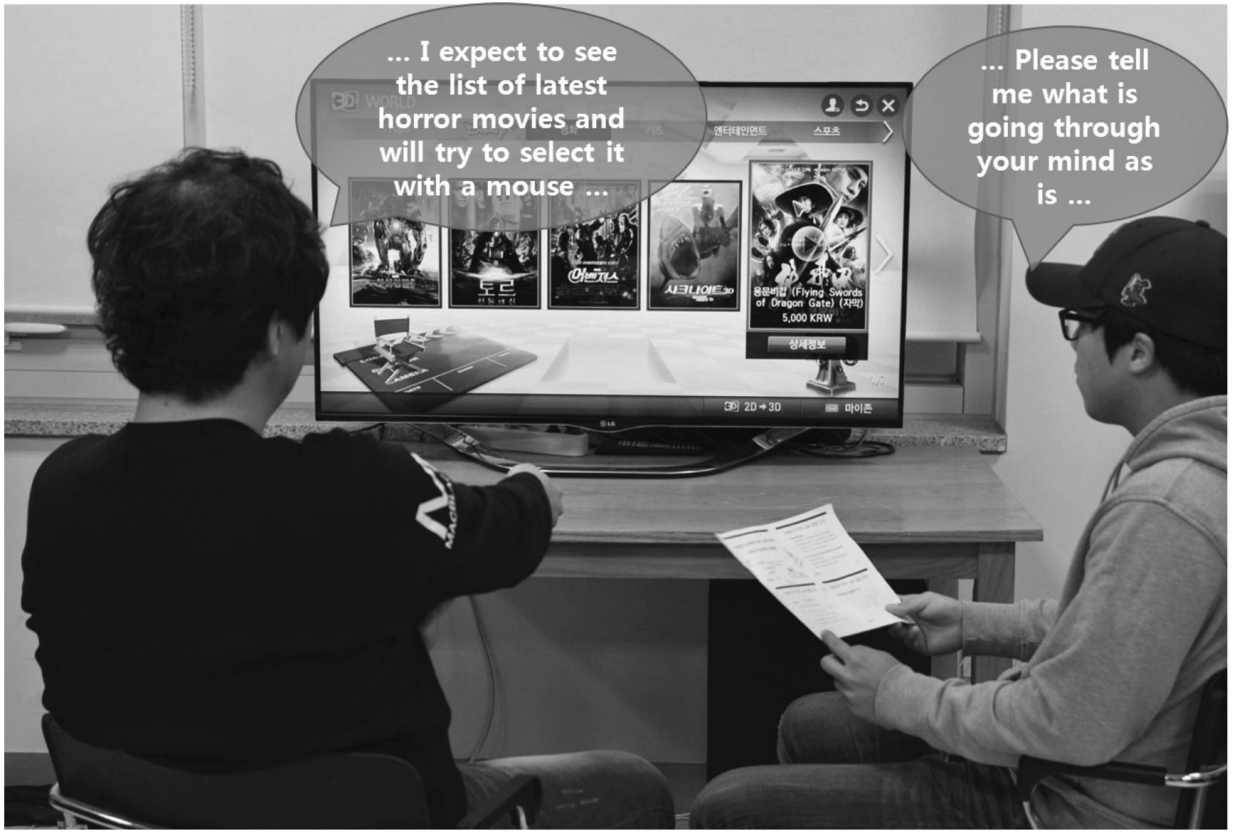
- Timing of analysis (e.g., throughout the application development stage: early, middle, late/after)
- Type and number of evaluators (e.g., several HCI experts vs. hundreds of domain users)
- Formality (e.g., controlled experiment or quick and informal assessment)
- Place of the evaluation (laboratory vs. field testing)

Focus Interview/Enactment/Observation Study:

- One of the easiest and most 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 involve actual system usage, or a paper/digital mock-up can be used for enacted scenarios. Mock-ups offer a tangible feel early in development, but may lack some interactive features.
- Wizard of Oz testing, with a human simulating system responses, is used when features are not implemented.
- The interview is often focused on particular user groups (e.g., elderly) or on the features of the system/interface (e.g., information layout) to save time.
- A particular interviewing technique, *The cognitive walkthrough* where the subject talks through their thought process. It's designed to identify gaps between the system's interaction model and the user's understanding. This method is suitable for evaluating the early stages of design, such as interaction modeling or interface selection, rather than specific interface design.

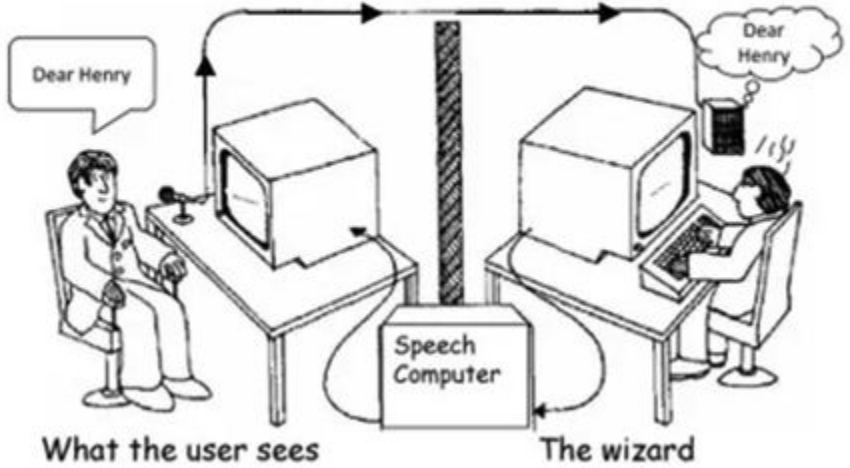


Interviewing a subject upon simulating the usage of the interface with a mock-up.



A cognitive walkthrough with the interviewer.

Wizard of Oz testing – The listening type writer IBM 1984



Focus Interview/Enactment/Observation Study:

- 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 flaws and interface-design-related bugs.

- 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

- The interview is free form, easy to administer but not structured or comprehensive.

Evaluators/size	Actual users/medium sized (10–15)		
Type of evaluators	Focused (e.g., by expertise, age group, gender)		
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, labeling, layout, etc.)	Simulation Actual system

Expert Heuristic Evaluation:

Very similar to the interview method. The difference is that the evaluators are HCI experts and that the analysis is carried out against a pre prepared HCI guideline. Nielsen's ten general UI heuristics

- **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 that 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 that 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 error 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, be focused on the user's task, list concrete steps to be carried out, and not be too large

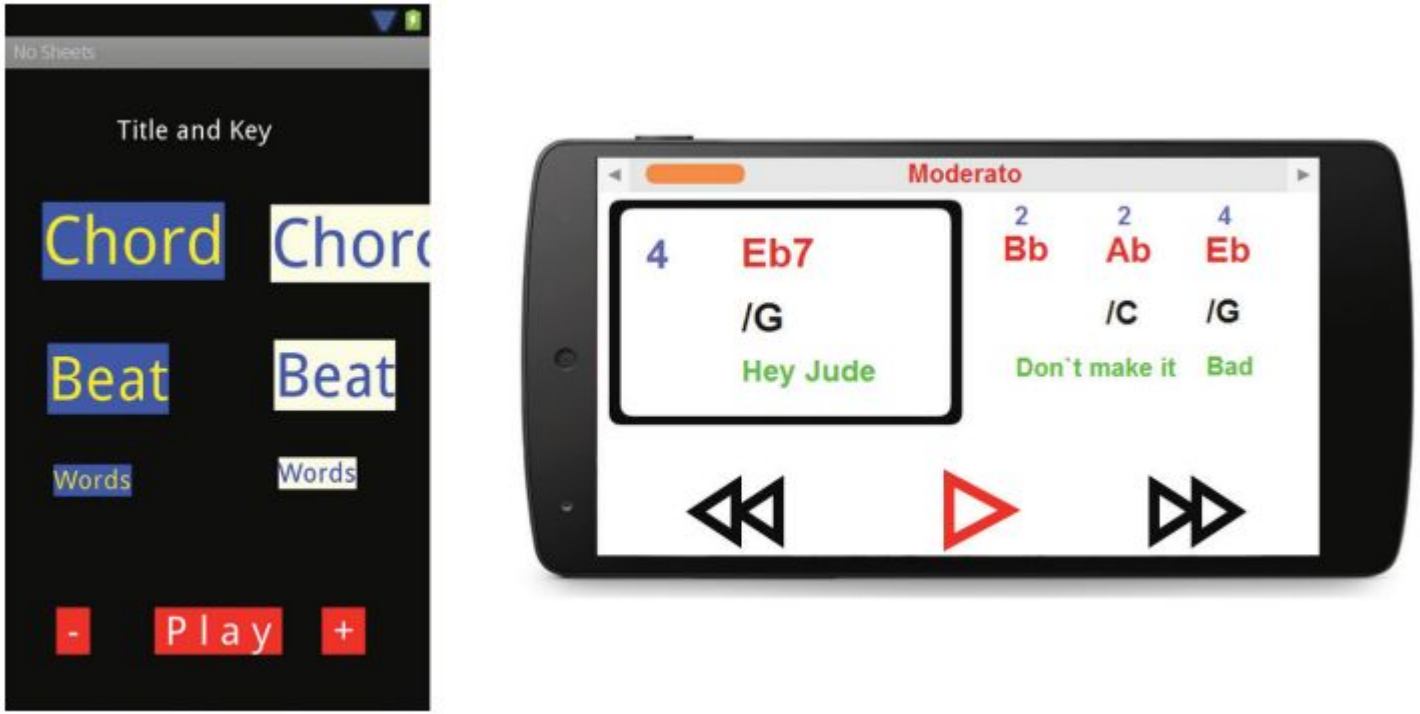
It is one of the most popular methods of UI evaluation because it is quick and dirty and relatively cost effective. Only a few (typically three to five) UI and domain experts are typically brought in to evaluate the UI implementation in the late stage of the development or even against a finished product.

The disadvantage of the expert review is that the feedback from the user is absent,

HEURISTIC	SPECIFICS (EXAMPLES)	EVALUATION RESULTS (PARTIAL)
System status	Does the user understand what is going on as the song is played (e.g., part of the song is being played, current operation)?	While playing, the tempo and whether it is being played, fast-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 and review are not familiar.
Interaction/ Contents model	Are all the essential functionalities available for this application? Are the necessary functions accessible and is information displayed at different interaction points?	Tempo control, fast-forward, and review are not possible during play. Information per measures is needed.
Ergonomic consideration/User characteristics/ Operating environment	Assess readability, color contrast, and 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: conveying the beat (beat number, sound), setting the tempo, selecting the song, etc.	Beat sound is too high pitched. 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 or designed such that it minimizes 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.

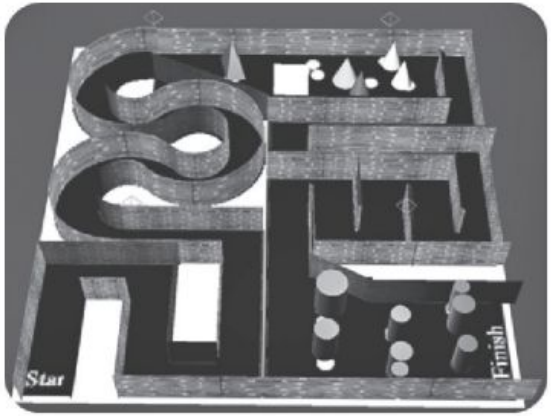
Evaluators/size	HCI experts/small sized (3–5)		
Type of evaluators	Focused (experts on application-specific HCI rules, corporate-specific design style, user ergonomics, etc.), interface consistency		
Formality	Usually informal (not controlled experiment)		
Timing and objectives	STAGE	OBJECTIVE	ENACTMENT 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

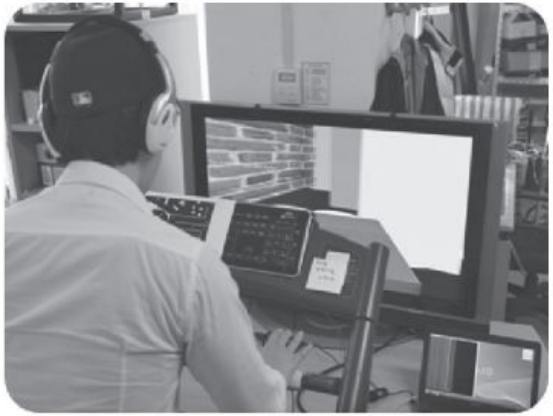


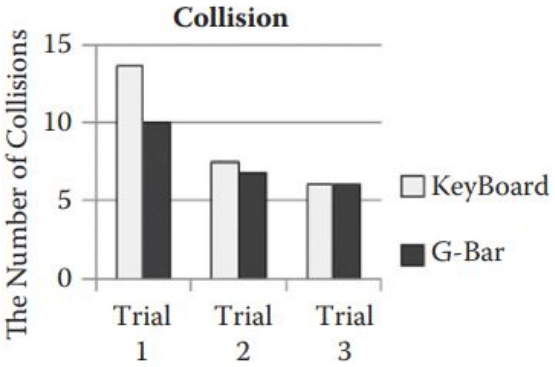
The initial (left) and redesigned (right) “play” activity/layer for No Sheets: The new design after evaluation uses a landscape mode and fewer primary colors. The icons for fast-forward and review are changed to the conventional style, and the current tempo is shown on top.

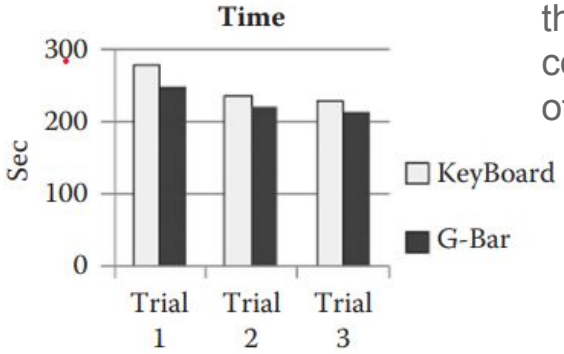
Measurement:

- Measurement methods aim to quantify interaction/interface design through task performance scores.
- Quantitative indicators include task completion time, score, and errors produced in unit time.
- Representative tasks, like logging into a mobile game or setting tempo in a music application, are used for measurement.
- Comparison with a nominal/reference case is essential for meaningful task-performance assessment.
- Statistical analysis is applied to identify significant differences between nominal and new design measurements.
- To minimize bias, a relatively small but homogeneous subject pool is recommended for task-performance measurement.



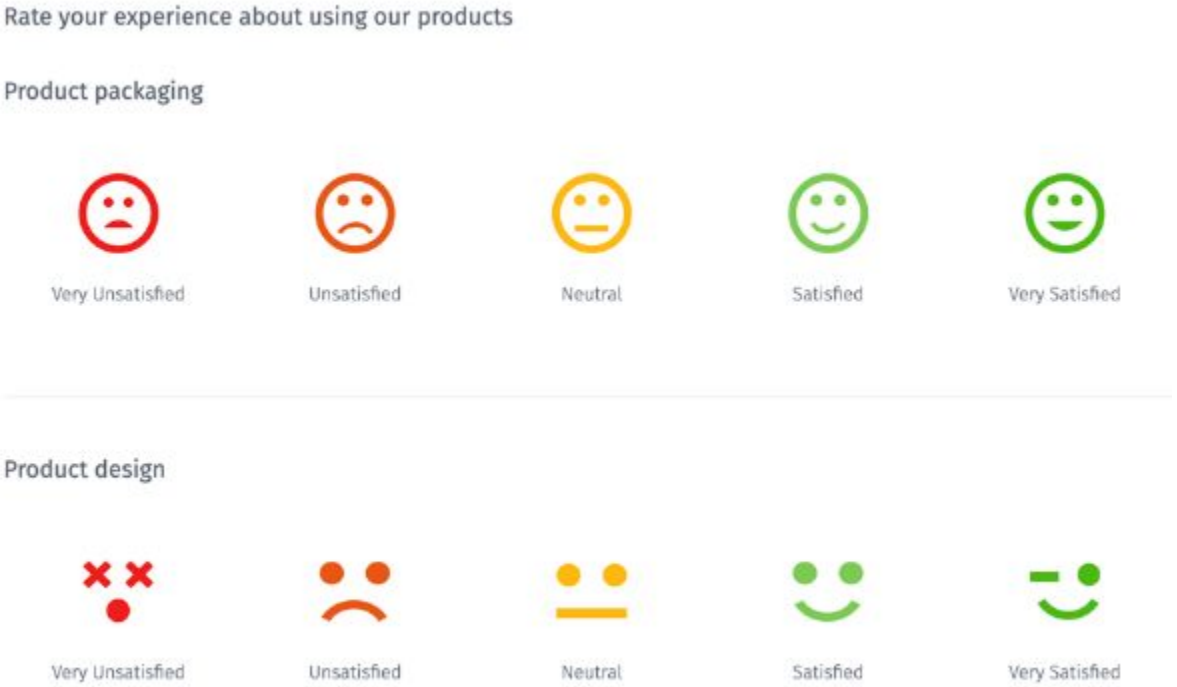






A case of a task-performance measurement: (1) nominal: a game interface using a keyboard, and (2) new: a game interface using a new controller. Task completion time for navigating a maze is measured using the respective interfaces and then compared to indirectly assess the ease of interaction

- Surveys provide numerical scores for aspects of usability and user experience based on perception.
- User-perception qualities are subjective and prone to bias, but measures like using a large subject pool and odd-leveled answer scales help mitigate bias.
- Survey questions are designed for clarity and understanding, often utilizing the Likert scale.
- Survey results, despite being numerical, are considered qualitative due to their focus on user perception.
- Comparative surveys against a nominal case are recommended for meaningful analysis.
- Both task-performance and survey experiments can be conducted over an extended period, especially for assessing memory performance and learning ease.



Example of Likert's Scale.

Minimize the number of questions	Too many questions results in fatigue and hence unreliable responses.
Use an odd-level scale of five or seven (or Likert Scale)	Research has shown odd answer levels with mid value with five or seven levels produces the best results.
Use consistent polarity	Negative responses correspond to Level 1 and positive to Level 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 their 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.

Guidelines for good survey.

- Ideal to conduct usage tests for finished products at the actual place of usage (office, home, street). Practical difficulties often lead to testing in a controlled laboratory environment with a homogeneous subject pool.

- Increasing popularity with smartphones; apps collect user interaction data for analysis in batch processes. Environmental biases exist but can be mitigated by a large subject pool, with comparable results to controlled laboratory studies depending on application nature.

- Meticulous planning needed for fair and bias-free measurement experiments. Includes recruitment, screening, pretraining, compensation, consent, variable selection, and appropriate statistical analysis.

- Detailed Design of Experiment required for measurement method reliability.Specifics beyond the book's scope; refer to related literature for more information.

- Despite higher reliability, substantial effort is necessary to prepare and administer measurement interface evaluation methods.

Safety and Ethics in Evaluation:

- Though safety problems rarely occur, precautions are still needed.
- For example, even interviews can become long and time consuming, causing the subject to feel much fatigue. Some seemingly harmless tasks may bring about unexpected harmful effects, both physically and mentally.
- Evaluations must be conducted on volunteers who have signed consent forms. Even with signed consents, the subjects have the right to discontinue the evaluation task at 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), which reviews the proposed evaluative experiments to ascertain safety and the rights of the subjects. It is best to consult or obtain permission from the IRB when there is even a small doubt of some kind of effect to the subjects during the experiments.