Evaluation

Introduction

- Designers can become so entranced with their creations that they may fail to evaluate them adequately.
- Experienced designers have attained the wisdom and humility to know that extensive testing is a necessity.
- Even if the design began with HCI considerations … → Still holes
  - Design is an iterative process anyway …
  - Priorities change
- Or the target software/system was a legacy ...

Evaluation Criteria: Usability

- Usability: usability refers to the ease of use and learnability of the user interface

- Quantitative Measures
  - Often involves task performance measurements.
  - 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.
  - However, assessment of a given new interface is better made in a comparative fashion against some nominal or conventional interface.

Evaluation Criteria: Usability

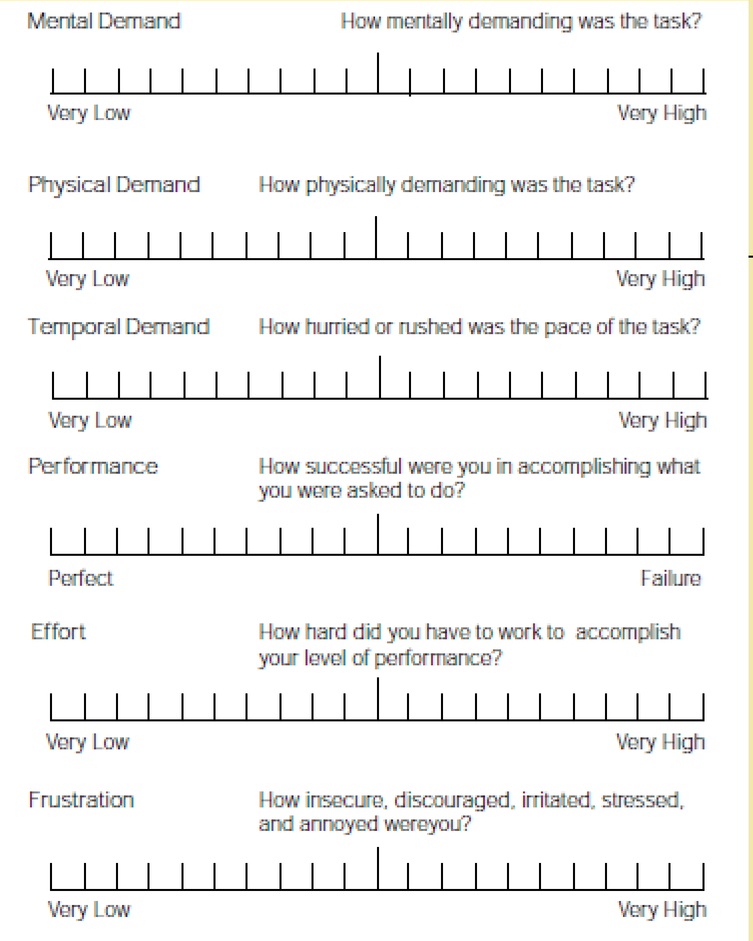
- Popular choices of such measure are
  - Task completion time
  - Task completion amount in a unit time (e.g. score)
  - Task error rate
  - For example, suppose we would like to test a new motion based interface for a smart phone game. We could have a pool of subjects play the game, using both the 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.

Evaluation Criteria: 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.
- Learnability generally involves much more biasing factors such as 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.

Evaluation Criteria

- Qualitative
  - To complement the shortcomings of the quantitative evaluation, **qualitative evaluations**
  - In most cases, qualitative 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.



Mental Demand How mentally demanding was the task?

Very Low Very High

Physical Demand How physically demanding was the task?

Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect Failure

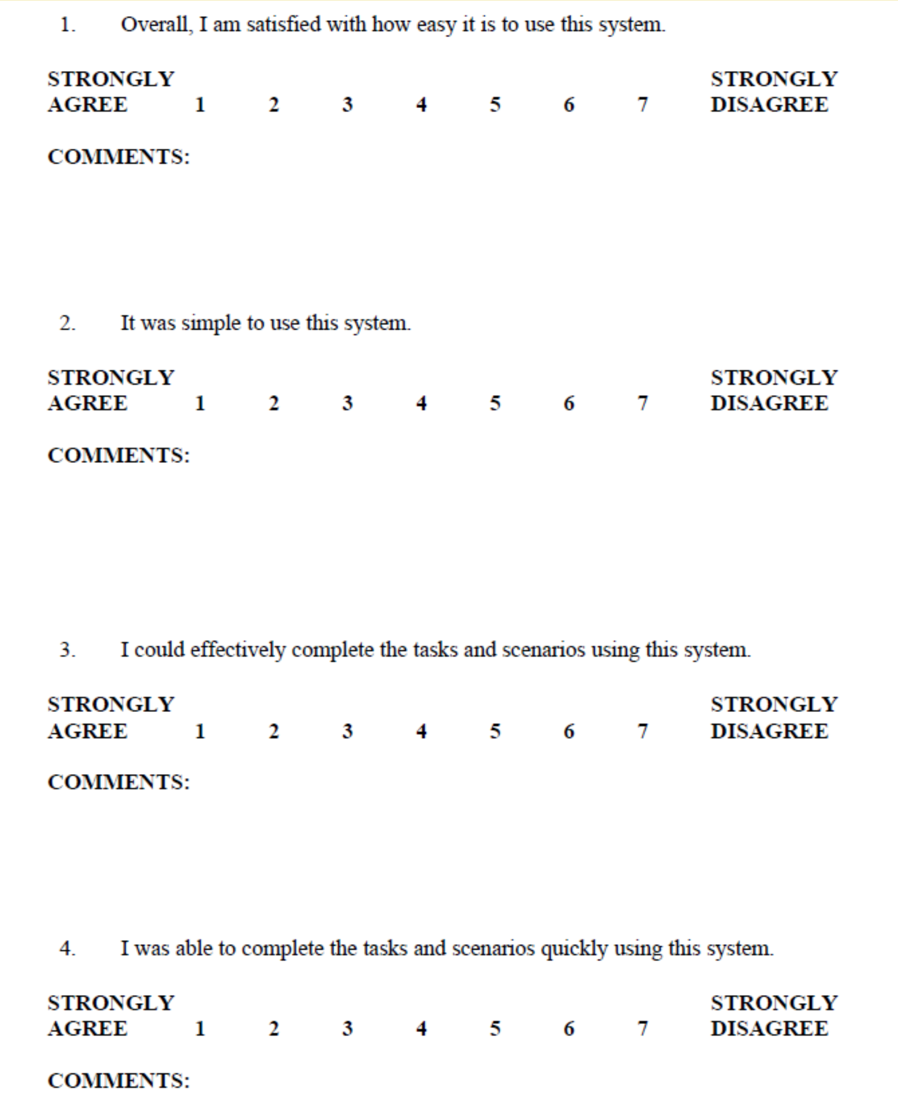
Effort How hard did you have to work to accomplish your level of performance?

Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low Very High

NASA TLX (Task Load Index) is one of the often used semi-standard questionnaires for this purpose [NASA]. NASA Task Load Index method assess work load on 7 point scales. 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 1 2 3 4 5 6 7 STRONGLY DISAGREE

COMMENTS:

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 this 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:

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

Evaluation Criteria: UX

- There is no precise definition for UX. It is generally accepted that user experience is total in the sense that it is not just about the interface but also 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 user or anticipated use of the application (through the given interface) [ISO 9241-210].
- Such affective response is very much dependent on the context of use.
- Thus UX evaluation involves a more comprehensive assessment on emotional response, under variety of usage contexts and across a family of products/applications/interfaces.
- A distinction can be made between usability methods hat have the objective of improving human performance, and user experience methods that have the objective of improving user satisfaction with achieving both pragmatic and hedonic goals [Bevan].
- Note that the notion of UX encompasses usability, that is, "usually" high UX translates to high usability and high emotional attachment.

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),
  - Place of evaluation (laboratory vs. in-situ field testing).

Focus Interview / Enactment / Observation Study

- One of the easiest and straightforward evaluation methods
- **Interview** the actual/potential users and **observe their interaction behavior** either with the finished product or through a simulated run.
- Simple question-&-answer form, and involves 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 test drive.
  - Simple paper/digital mock-up is used so that a particular usage scenario can be 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 in question (vs. a mere rough sketch), at an early stage of the development, important interactivity may not have been implemented. 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 system usage or simulation are recorded or video-taped for more detailed analysis

Focus Interview / Enactment / Observation Study

- 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 relatively the 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 flow and interface design related “bugs.”

Interface prototyping

Focus Interview / Enactment / Observation Study

- 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 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.

Summary

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

Expert/Heuristic Reviews

- 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 (almost like the “design” principles) or more specific, 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.

- While informal demos to colleagues or customers can provide some useful feedback, more formal expert reviews have proven to be effective

Expert/Heuristic Reviews

- One of the most popular methods of UI evaluation because it is quick and dirty and relatively cost effective (involving only few UI experts).
- A few (typically 3~5) UI and domain experts are 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.
  - Even experienced expert reviewers have great difficulty knowing how typical users, especially first-time users will really behave.
- The small sized evaluator pool is compensated by their expertise.

Expert/Heuristic Reviews

- Expert reviews can be scheduled at several points in the development process when experts are available and when the design team is ready for feedback.

- Different experts tend to find different problems in an interface, so 3-5 expert reviewers can be highly productive, as can complementary usability testing.

- Expert reviews entail one-half day to one week effort, although a lengthy training period may sometimes be required to explain the task domain or operational procedures

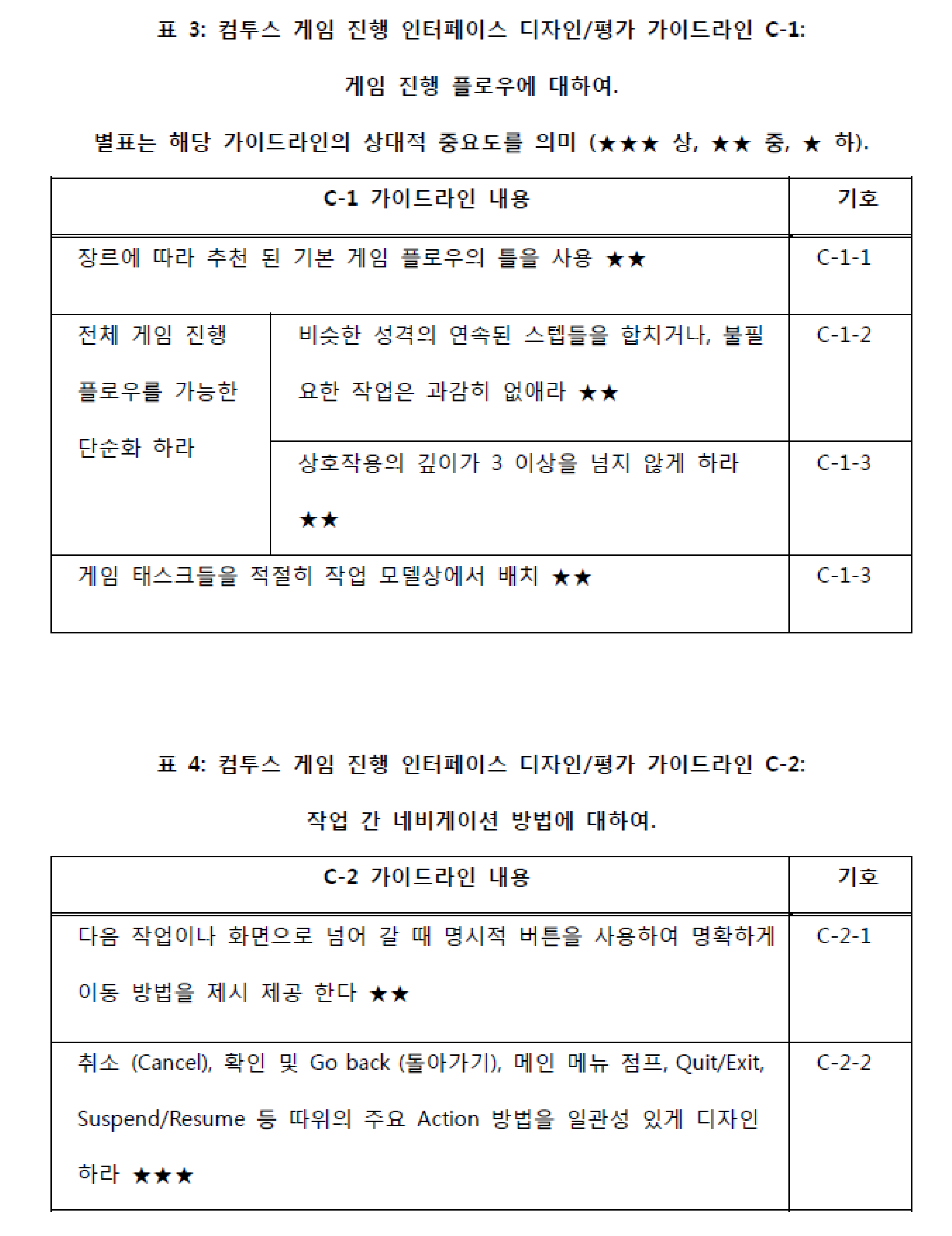


표 3: 컴투스 게임 진행 인터페이스 디자인/평가 가이드라인 C-1:

게임 진행 플로우에 대하여.

별표는 해당 가이드라인의 상대적 중요도를 의미 (★★★ 상, ★★ 중, ★ 하).

C-1 가이드라인 내용		기호
장르에 따라 추천 된 기본 게임 플로우의 틀을 사용 ★★		C-1-1
전체 게임 진행 플로우를 가능한 단순화 하라	비슷한 성격의 연속된 스텝들을 합치거나, 불필요한 작업은 과감히 없애라 ★★	C-1-2
	상호작용의 깊이가 3 이상을 넘지 않게 하라 ★★	C-1-3
게임 태스크들을 적절히 작업 모델상에서 배치 ★★		C-1-3

표 4: 컴투스 게임 진행 인터페이스 디자인/평가 가이드라인 C-2:

작업 간 네비게이션 방법에 대하여.

C-2 가이드라인 내용	기호
다음 작업이나 화면으로 넘어 갈 때 명시적 버튼을 사용하여 명확하게 이동 방법을 제시 제공 한다 ★★	C-2-1
취소 (Cancel), 확인 및 Go back (돌아가기), 메인 메뉴 점프, Quit/Exit, Suspend/Resume 등 따위의 주요 Action 방법을 일관성 있게 디자인 하라 ★★★	C-2-2

표 5: 컴투스 게임 진행 인터페이스 디자인/평가 가이드라인 C-3:

화면/위젯 레이아웃 방법에 대하여.

C-3 가이드라인 내용	기호
게임의 상태 (점수, 주 아이템 소유, 게임 시간/프로그레스 등) 의 정보는 되도록 화면 상단에 배치 ★	C-3-1
게임의 조작 인터페이스나 주 기능들은 오른쪽 아래 배치 (두 손이 필요한 경우 왼쪽 아래에도 배치 가능) ★★	C-3-2
Pop up 위젯에서 되도록 아래의 가이드라인을 지킨다: (1) Exit 버튼: 오른쪽 위 (혹은 아래), (2) 확인 버튼: 중앙 아래, (3) 제목: 중앙 위 ★★	C-3-3
Pop up 위젯을 사용 할 때와 새 화면으로 전환 해야 할 때를 구분. Pop up 의 경우 배경 화면 클릭을 통해 확인/돌아가기가 가능하게 하고, 화면 전환의 경우 돌아가기 버튼을 명확하게 표시 ★	C-3-4

표 6: 컴투스 게임 진행 인터페이스 디자인/평가 가이드라인 C-4:

인간공학적 UI 객체 디자인에 대하여.

C-4 가이드라인 내용	기호
버튼/아이콘/라벨/텍스트 등 선택 인터페이스의 크기가 손가락 끝 크기 및 사용자 일반 시력을 고려 하여야 한다 ★★★	C-4-1
버튼, 아이콘등과 배경 사이, 라벨과 배경 사이, Pop up 과 배경 화면 사이 등 충분한 Contrast 나 정보/내용 복잡도/밀도의 조정에 의하여 명확하게 분리 되어 인지 될 수 있게 디자인. 필요 없이 Highlight 하지 말 것 ★★★	C-4-2
C-4-3: 라벨이나 아이콘 그림들이 가능한 이해 가능 해야 하며 이해를 돕기 위해 라벨과 아이콘을 모두 사용 하는 것을 고려 ★★★	C-4-3

표 7: 컴투스 게임 진행 인터페이스 디자인/평가 가이드라인 C-5:

인터페이스 일관성에 대하여.

C-5 가이드라인 내용	기호
주어진 장르에 대해서 비슷한 작업 모형을 채택 하여 Flow Consistency를 유지 ★★	C-5-1
Image work 애셋을 활용 하여 비슷한 Look and Feel Consistency를 성취 유지. 게임의 특성에 기반 하여 모든 화면의 Look and Feel을 통일 ★★	C-5-2
버튼/아이콘 Location Consistency를 되도록 유지 ★★	C-5-3

표 8: 컴투스 게임 진행 인터페이스 디자인/평가 가이드라인 C-6:

광고/컴투스홍보에 대하여.

C-6 가이드라인 내용	기호
광고가 주요 정보를 가리지 않도록 디자인 ★★	C-6-1
게임 플로우에 광고를 너무 과대하게 삽입 하지 않도록 한다 ★★	C-6-2

표 9: 컴투스 게임 진행 인터페이스 디자인/평가 기타 가이드라인 C-7:

게임 액션 인터페이스에 대하여.

C-7 가이드라인 내용	기호
게임 액션/조작을 위한 인터페이스는 체험성이나 동작의 은유적 매핑 보다 간편성과 신속성을 우선시 하라 ★★	C-7-1
장르 사이에 게임 액션/조작을 위한 인터페이스의 일관성을 지켜라 (Action Consistency) ★★	C-7-2 (C-5-4)

Summary

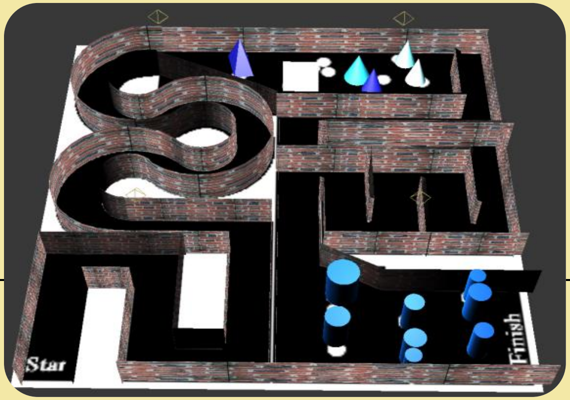
<table>
<tr><td><b>Evaluators / Size</b></td><td colspan="3"><b>HCI Experts / Small sized (3~5)</b></td></tr>
<tr><td><b>Type of evaluators</b></td><td colspan="3">Focused (Experts on Application specific HCI rules, Corporate specific design style, User ergonomics, etc.), Interface consistency</td></tr>
<tr><td><b>Formality</b></td><td colspan="3">Usually informal (not controlled experiment)</td></tr>
<tr><td rowspan="3"><b>Timing and Objectives</b></td><td>Stage</td><td>Objective</td><td>Enactment Method</td></tr>
<tr><td>Middle</td><td>Interface selection</td><td>Scenarios
Storyboards
Interaction Model</td></tr>
<tr><td>Late/After</td><td>Interface design issues (look and feel such as aesthetics, color, contrast, font size, icon location, labeling, layout, etc.)</td><td>Simulation
Actual system</td></tr>
<tr><td colspan="4"><b>Easy and quick, but prior heuristics assumed to exist and no actual user feedback reflected</b></td></tr>
</table>

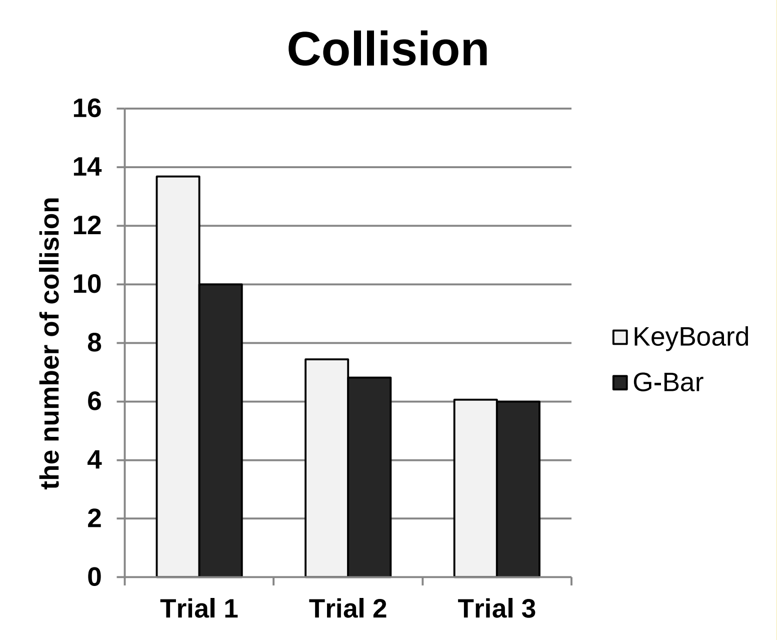
Measurement

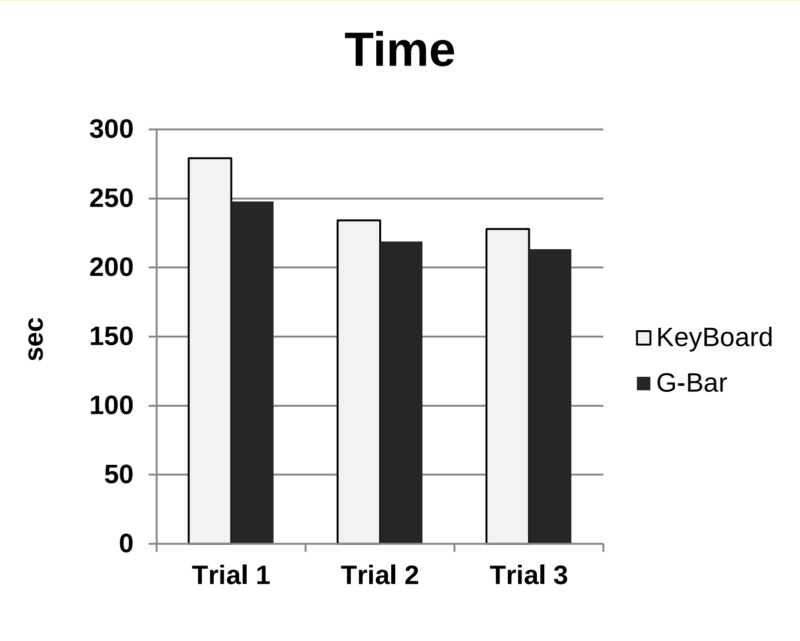
- Measurement methods attempt to indirectly quantify the goodness of the interaction/interface design with **a score** through representative task performance (quantitative) or quantified answers from carefully prepared subjective surveys (qualitative).

- Typical indicators for quantitative task performance are 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, for “No Sheets” similarly 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 difference between the two measurements.







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.

Surveys

- 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 so that there always exist the middle level answer,
  - carefully wording and explaining the survey question for clarity and understanding

- 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.

Survey Guidelines

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.

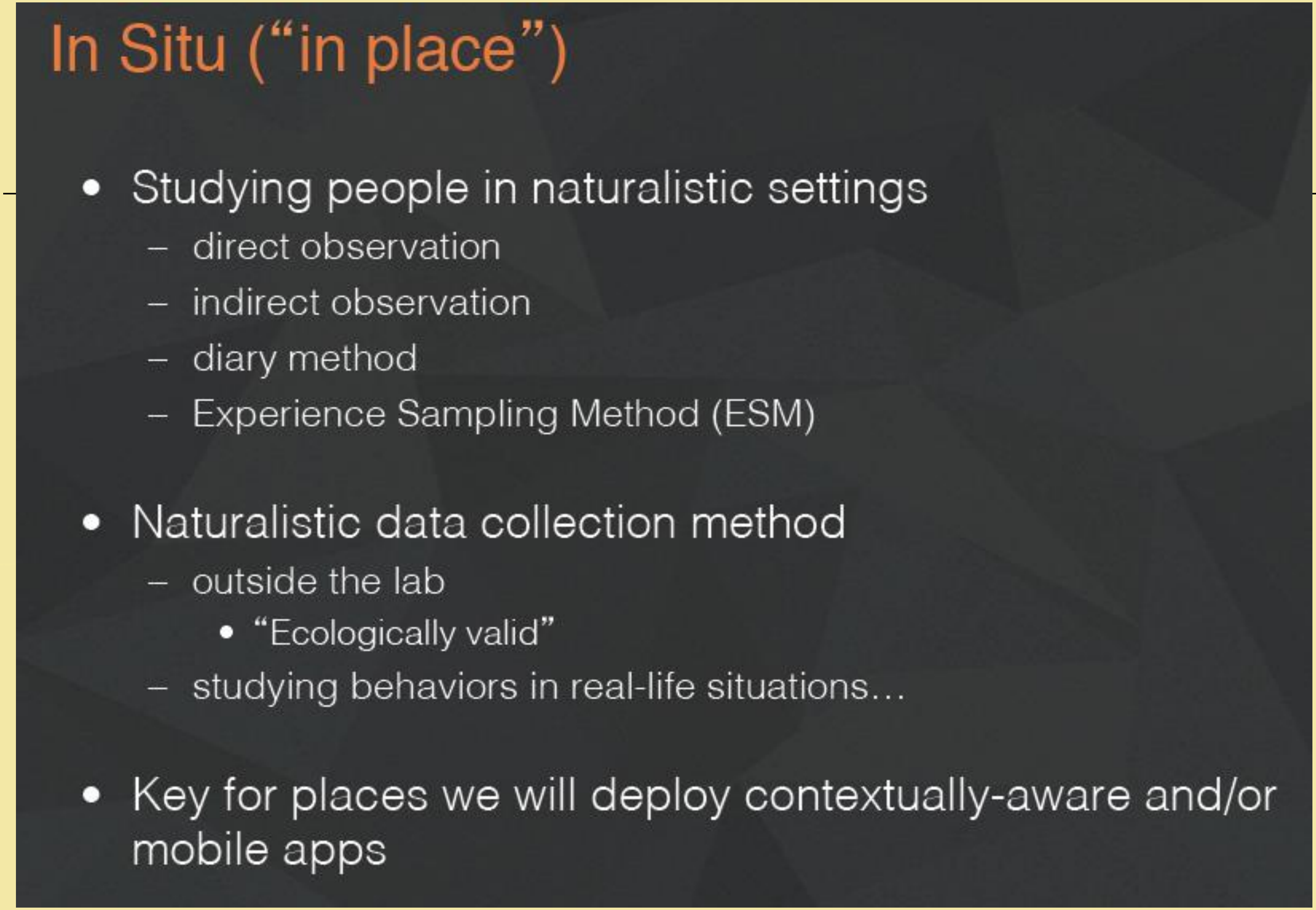
Measurement: Other considerations

- 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.). H
  - However, 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.

Experience Sampling (Method): *Evaluation during Active Usage*

- With the advent of the smart phones and their ubiquity, the in-situ field testing is gaining great popularity.
- Applications can collect user interaction information in the background upon particular interaction events and be analyzed in a batch process.
- While the same danger exist with respect to the environmental biases, they are often mitigated by the high number of subjects (e.g. users of smart phones and apps).
- Research has shown that there is very little difference in the analysis/evaluation results between the controlled laboratory studies and in-situ field studies [Ref].

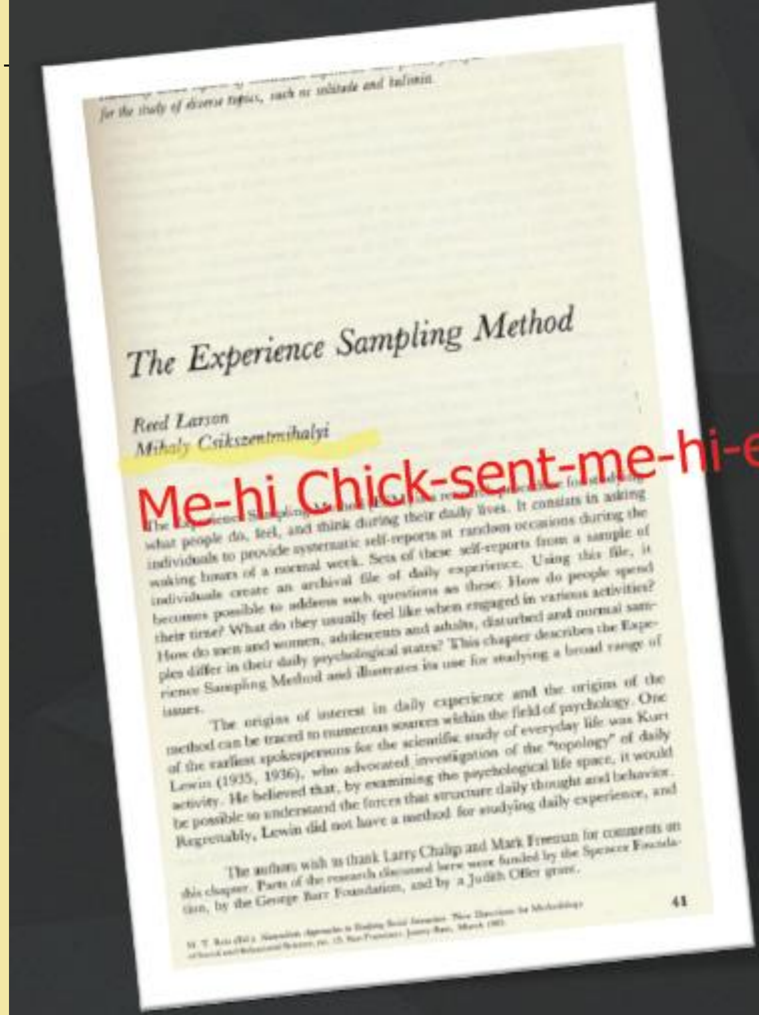


In Situ ("in place")

- Studying people in naturalistic settings
  - direct observation
  - indirect observation
  - diary method
  - Experience Sampling Method (ESM)
- Naturalistic data collection method
  - outside the lab
    - "Ecologically valid"
  - studying behaviors in real-life situations…
- Key for places we will deploy contextually-aware and/or mobile apps

James Landay, University of Washington

Experience Sampling Method (ESM)



... for the study of diverse topics, such as solitude and [illegible]

The Experience Sampling Method

Reed Larson
Mihaly Csikszentmihalyi

The Experience Sampling Method (ESM) is a research procedure for studying what people do, feel, and think during their daily lives. It consists in asking individuals to provide systematic self-reports at random occasions during the waking hours of a normal week. Sets of these self-reports from a sample of individuals create an archival file of daily experience. Using this file, it becomes possible to address such questions as these: How do people spend their time? What do they usually feel like when engaged in various activities? How do men and women, adolescents and adults, disturbed and normal samples differ in their daily psychological states? This chapter describes the Experience Sampling Method and illustrates its use for studying a broad range of issues.

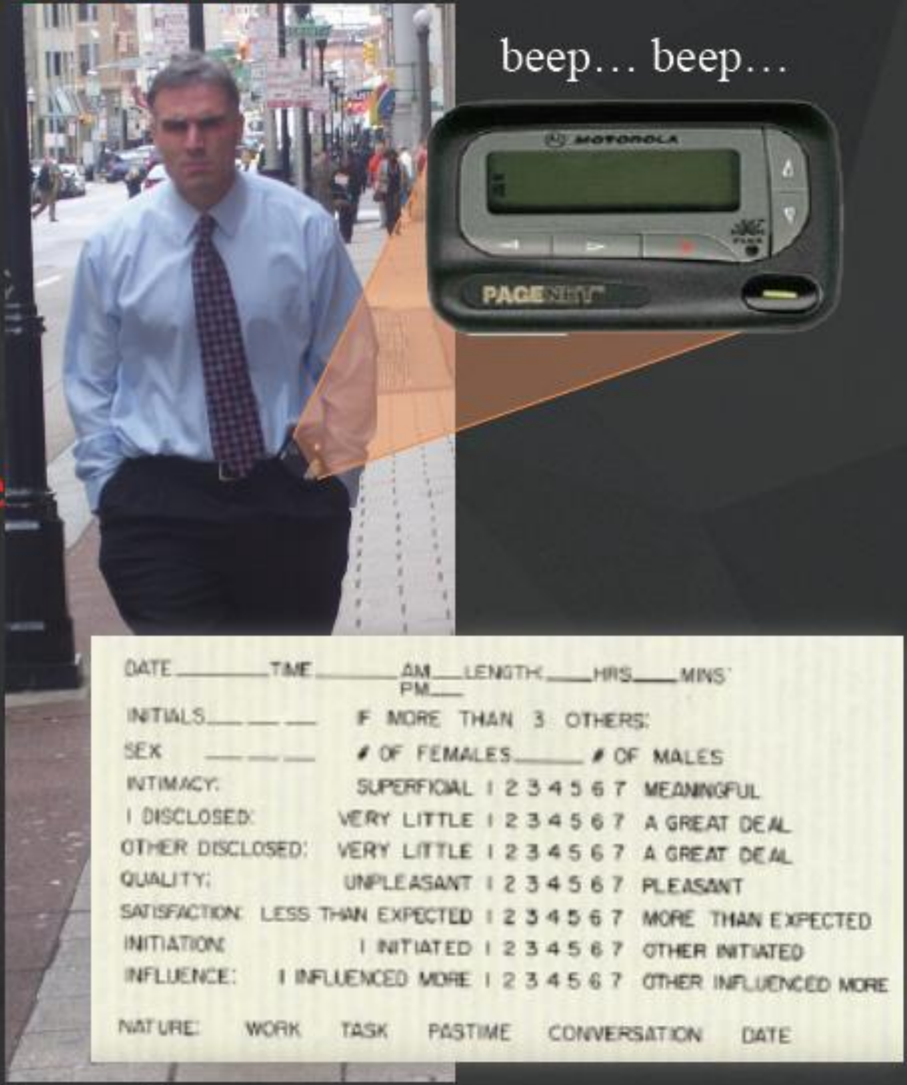
The origins of interest in daily experience and the origins of the method can be traced to numerous sources within the field of psychology. One of the earliest spokespersons for the scientific study of everyday life was Kurt Lewin (1935, 1936), who advocated investigation of the "topology" of daily activity. He believed that, by examining the psychological life space, it would be possible to understand the forces that structure daily thought and behavior. Regrettably, Lewin did not have a method for studying daily experience, and

The authors wish to thank Larry Chalip and Mark Freeman for comments on this chapter. Parts of the research discussed here were funded by the Spencer Foundation, by the George Barr Foundation, and by a Judith Offer grant.

41

M. T. Reis (Ed.). Naturalistic Approaches to Studying Social Interaction. New Directions for Methodology of Social and Behavioral Science, no. 15. San Francisco: Jossey-Bass, March 1983.

Me-hi Chick-sent-me-hi-ee



beep… beep…

DATE ______ TIME ______ AM___ PM___ LENGTH: ___HRS___MINS

INITIALS ___ ___ ___ IF MORE THAN 3 OTHERS:

SEX ___ ___ ___ # OF FEMALES_____ # OF MALES

INTIMACY: SUPERFICIAL 1 2 3 4 5 6 7 MEANINGFUL

I DISCLOSED: VERY LITTLE 1 2 3 4 5 6 7 A GREAT DEAL

OTHER DISCLOSED: VERY LITTLE 1 2 3 4 5 6 7 A GREAT DEAL

QUALITY: UNPLEASANT 1 2 3 4 5 6 7 PLEASANT

SATISFACTION: LESS THAN EXPECTED 1 2 3 4 5 6 7 MORE THAN EXPECTED

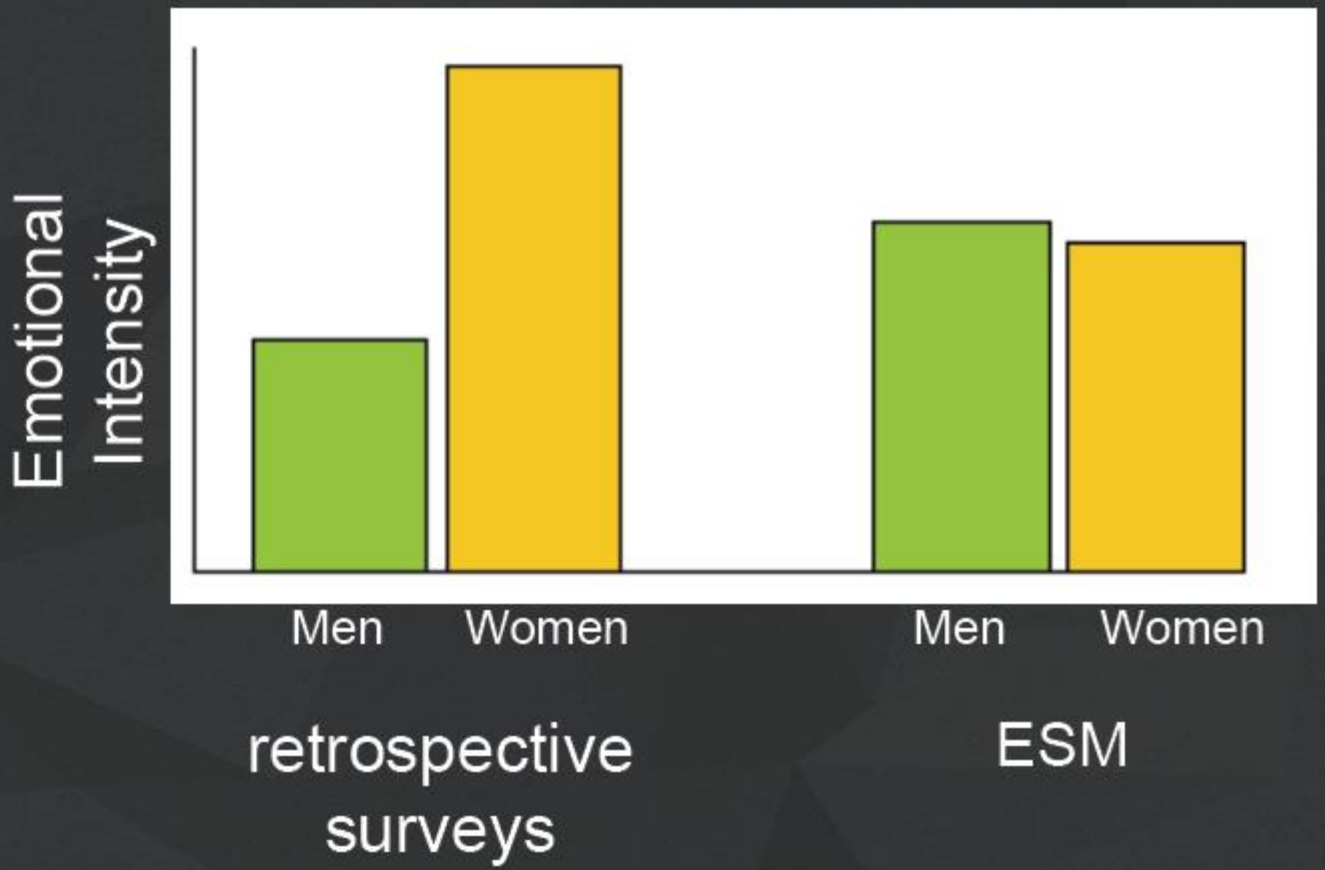
INITIATION: I INITIATED 1 2 3 4 5 6 7 OTHER INITIATED

INFLUENCE: I INFLUENCED MORE 1 2 3 4 5 6 7 OTHER INFLUENCED MORE

NATURE: WORK TASK PASTIME CONVERSATION DATE

Also called "signal-contingent" sampling...

Why is ESM Interesting?

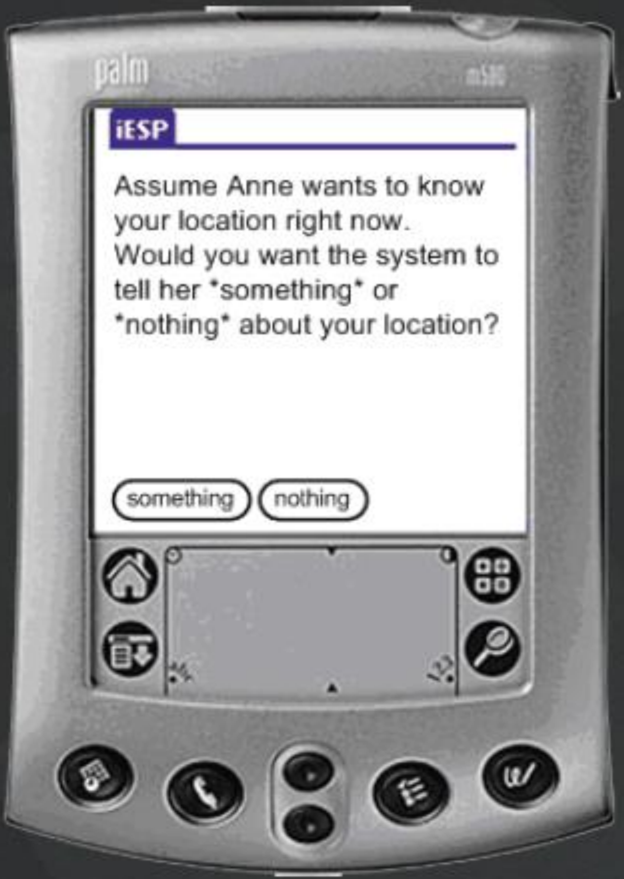


Barrett, *Cognition and Emotion*, 1998

Computerized ESM

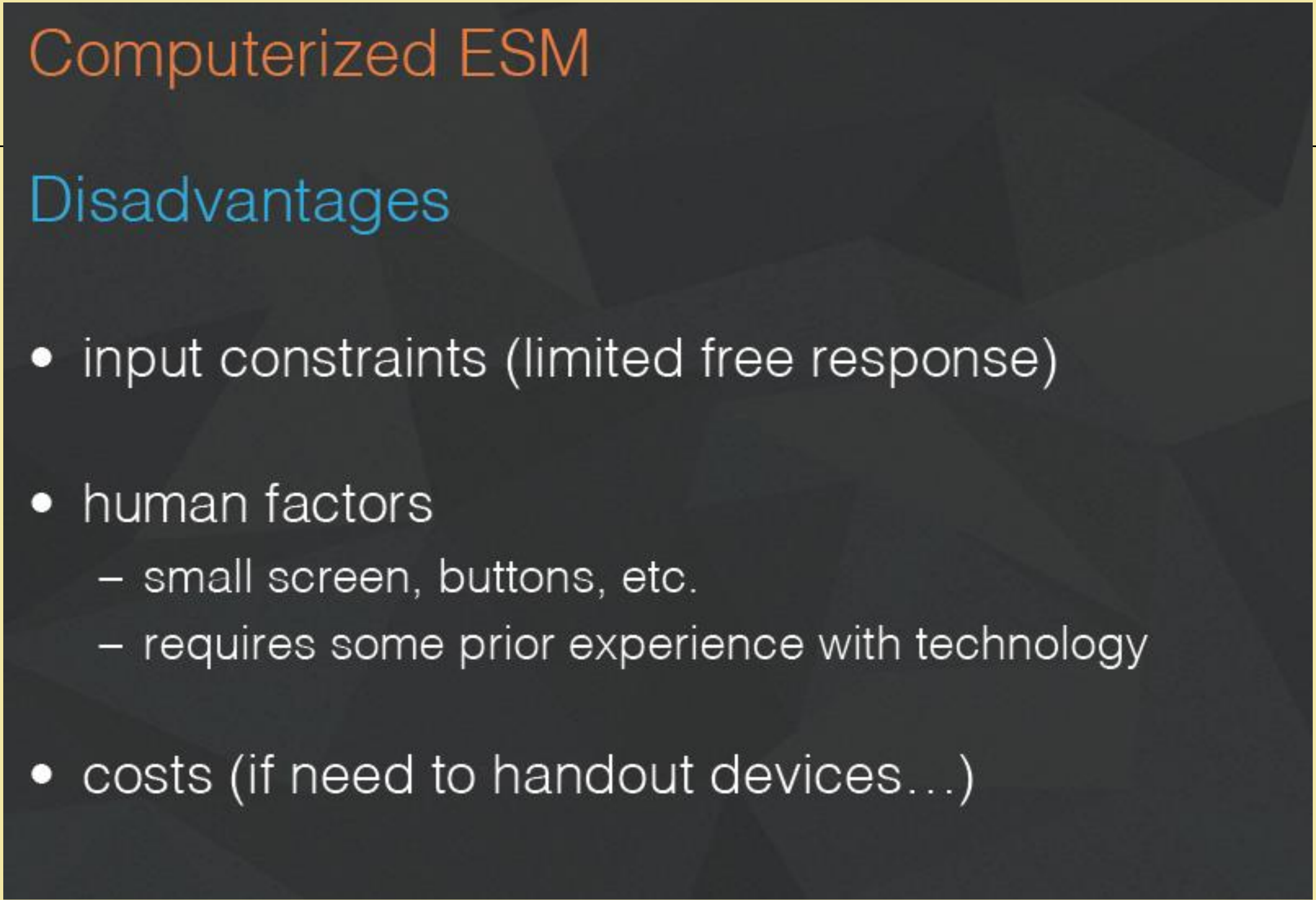
Advantages

- ensures compliance
- sophisticated presentation
  - conditionals
  - probabilities
  - “question pools”
- record reaction times
- data already in computer
  - reduces data entry error



iESP

Assume Anne wants to know your location right now. Would you want the system to tell her *something* or *nothing* about your location?



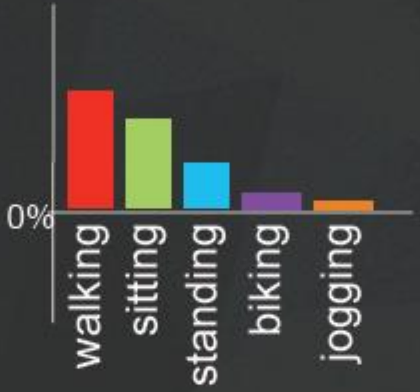
Computerized ESM

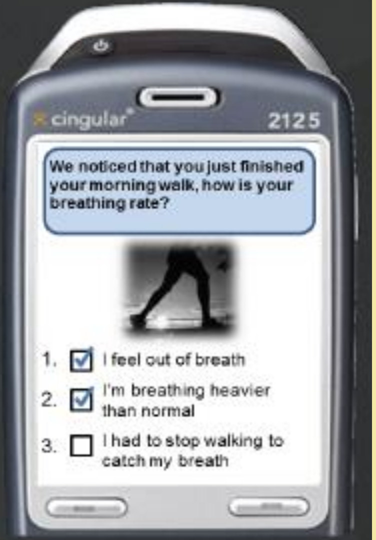
Disadvantages

- input constraints (limited free response)
- human factors
  - small screen, buttons, etc.
  - requires some prior experience with technology
- costs (if need to handout devices…)

Context-Triggered Sampling

- Use sensors to achieve targeted triggers
- Do not need to bug the customers as often
  - e.g., after a walk, in a certain place, etc.





We noticed that you just finished your morning walk, how is your breathing rate?

1. ☑ I feel out of breath
2. ☑ I'm breathing heavier than normal
3. □ I had to stop walking to catch my breath

Example Triggers

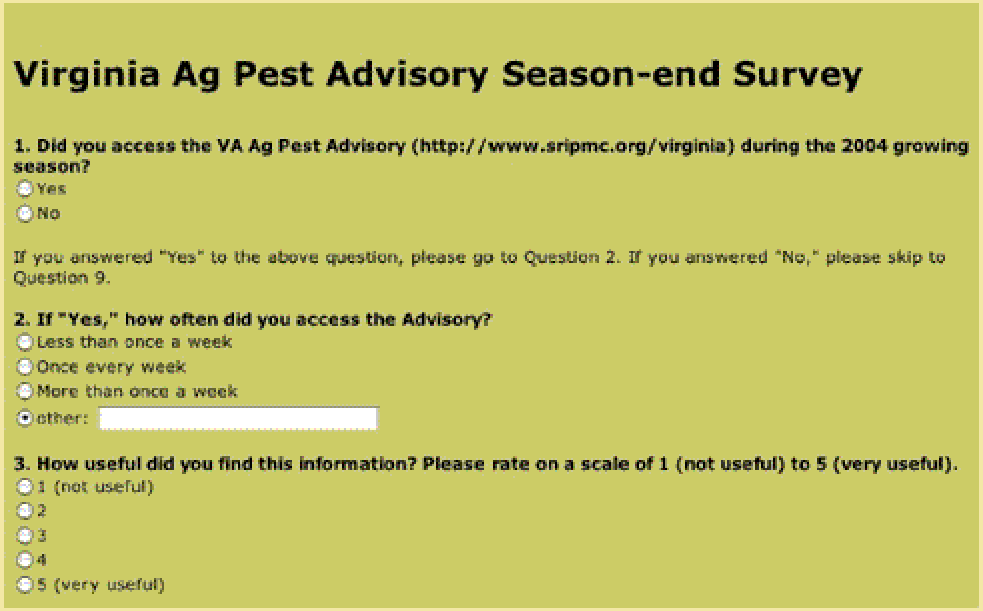
Activity == walking
DeviceIdle > 15 mins
Place.State == "Home"

Example Actions

SurveyAction
ScreenshotAction
LogAction

Online Surveys

- Online surveys avoid the cost of printing and the extra effort needed for distribution and collection of paper forms.
- Many people prefer to answer a brief survey displayed on a screen, instead of filling in and returning a printed form,
  - although there is a potential bias in the sample.



Virginia Ag Pest Advisory Season-end Survey

1. Did you access the VA Ag Pest Advisory (http://www.sripmc.org/virginia) during the 2004 growing season?

○ Yes
○ No

If you answered "Yes" to the above question, please go to Question 2. If you answered "No," please skip to Question 9.

2. If "Yes," how often did you access the Advisory?

○ Less than once a week
○ Once every week
○ More than once a week
● other:

3. How useful did you find this information? Please rate on a scale of 1 (not useful) to 5 (very useful).

○ 1 (not useful)
○ 2
○ 3
○ 4
○ 5 (very useful)

Summary

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 형식	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			

Usability Testing and Laboratories

- The emergence of usability testing and laboratories since the early 1980s
- Usability testing not only sped up many projects but that it produced dramatic cost savings.
- The movement towards usability testing stimulated the construction of usability laboratories.
- A typical modest usability lab would have two 10 by 10 foot areas, one for the participants to do their work and another, separated by a half-silvered mirror, for the testers and observers
- Participants should be chosen to represent the intended user communities, with attention to
  - background in computing, experience with the task, motivation, education, and ability with the natural language used in the interface.

Usability Testing and Laboratories

- Participation should always be voluntary, and informed consent should be obtained.

- Professional practice is to ask all subjects to read and sign a statement like this one:
  - I have freely volunteered to participate in this experiment.
  - I have been informed in advance what my task(s) will be and what procedures will be followed.
  - I have been given the opportunity to ask questions, and have had my questions answered to my satisfaction.
  - I am aware that I have the right to withdraw consent and to discontinue participation at any time, without prejudice to my future treatment.
  - My signature below may be taken as affirmation of all the above statements; it was given prior to my participation in this study.

Usability Testing and Laboratories

Videotaping participants performing tasks is often valuable for later review and for showing designers or managers the problems that users encounter.

Measurement Experiments

- Measurement experiments require a meticulous operational logistics
- To be as fair and bias free as possible,
  - Recruitment and screening of the subjects,
  - Pre-training of them,
  - Compensation and obtaining consent,
  - Choosing the right independent and dependent variables
  - Applying the right statistical analysis methods to the resulting data.

Carrying out Experiments:

- Where
  - In the lab
    - More controlled
    - No distractions
  - In the Field
    - More realistic
    - Discovery of realistic problems

	K/M	TS
	G1	G2

- What
  - Independent variables (what we want to control)
  - Dependent variables (what we want to observe)

e.g. Compare UI with keyboard/Mouse and UI with touch screen.
Independent var. → what you are controlling
Dependent var. → task completion time, answers to survey questions

E.g. 2 Factor 3 Levels

- Compare effects of UI style, age
- 2 Factors
  - UI style
    - Keyboard/Mouse
    - Touch screen
    - Voice driven
  - Age
    - Kid
    - Young
    - Old

How many combination? 3x3 = 9

	K/M	TS	V
K	G1	G4	G7
Y	G2	G5	G8
O	G3	G6	G9

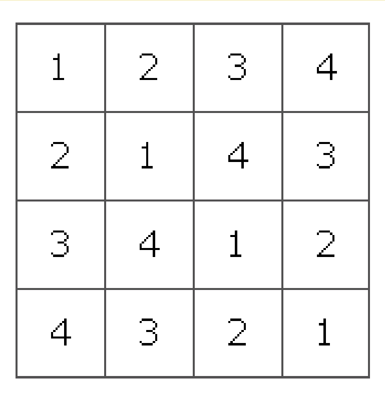
Experiment Design

- Repeated measure design(~ within group/subject)
  - A group/subject tries all combinations
    (They are their own "control" groups")
  - Ordering effects
    - Learning effect
    - Fatigue effect
    - Counter balance the experiment order (~ randomize)
  - Takes long time for subject (one subject tries all, need breaks …)

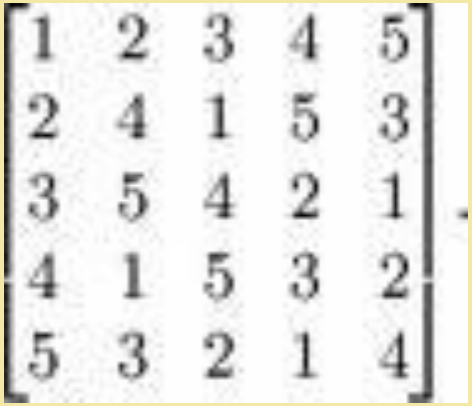
- Independent sample design (between group/subject)
  - Different subject pool for each group (costly)
  - Takes long time for experimenter
  - Difficult to normalize among groups
    - Need many subjects

Balanced Order

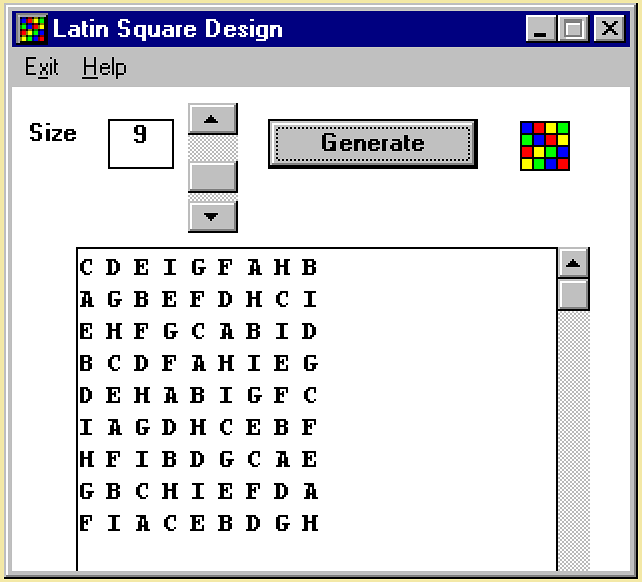
- A **Latin square** is an *n* × *n* table filled with *n* different symbols in such a way that each symbol occurs exactly once in each row and exactly once in each column.



1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1



$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 1 & 5 & 3 \\ 3 & 5 & 4 & 2 & 1 \\ 4 & 1 & 5 & 3 & 2 \\ 5 & 3 & 2 & 1 & 4 \end{bmatrix}.$$



Size 9

```
C D E I G F A H B
A G B E F D H C I
E H F G C A B I D
B C D F A H I E G
D E H A B I G F C
I A G D H C E B F
H F I B D G C A E
G B C H I E F D A
F I A C E B D G H
```

Experiment process

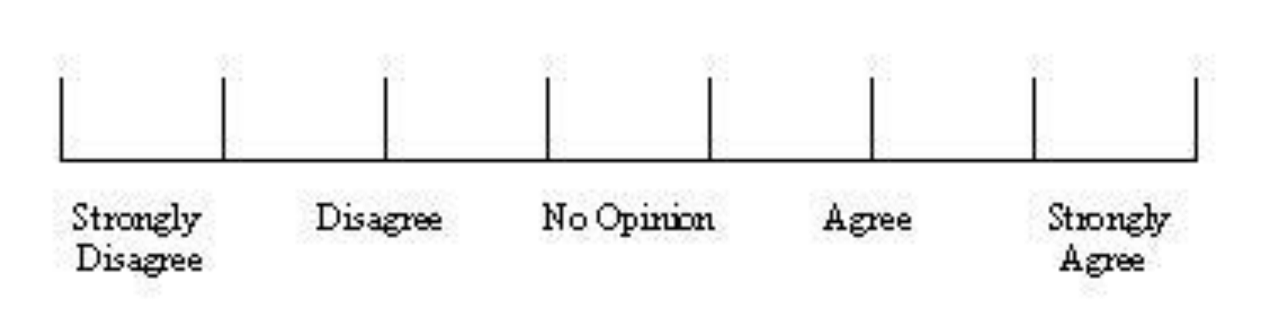
- Experiment is for “hypothesis testing”
  - Null hypothesis: Usually “factors do not matter”
  - Alternative hypothesis: the opposite
  - Which design to use?
  - How many subjects? → 30

- Accept/Reject null/alternative hypothesis
  - Based on analysis of stats (D.V.) of experiment
  - One tailed and two tailed test*
  - How do you do this analysis?

Measuring data

- Subjective Questionnaire (Survey)
  - Scaled questionnaire (Likert scale)
  - Semantic differential scaled questionnaire
  - Caution
    - Equal number of positives and negatives
      - Remove acquiescence effect
    - Make questions understandable
      - Do not put similar/redundant questions
      - Conduct pilot test and reconfigure the questions
    - Midpoint problem (5 or 7 level)
      - 5 or 7 (how discriminating the levels are?)
    - Polarity of scale
      - Ease of analysis
      - No confusion
    - Not too long! (Fatigue effect)

- Quantitative data
  - Task completion time (or some other performance data)
  - Error
  - Retention (Day 2 test) (above over time)



Semantic Differential Scale

For each pair of adjectives place a cross at the point between them which reflects the extent to which you believe the adjectives describe policemen

clean : : : : : : : : dirty

honest : : : : : : : : dishonest

kind : : : : : : : : cruel

helpful : : : : : : : : unhelpful

fair : : : : : : : : biassed

strong : : : : : : : : weak

foolish : : : : : : : : wise

energetic : : : : : : : : lazy

unreliable : : : : : : : : reliable

(Robson,1993)

Additionals

- Observation data
  - Recording behavior and analysis

- Interview / Debriefing
  - Informal / formal (Questionnaire in voice)
  - Verbal protocol / Think aloud / Cognitive walkthrough
  - Post event protocol
    - Bring back subjects and ask about video

Measurement Data Types

- Numerical
  - Interval scale: Numerical data itself
  - Ordinal scale: Rank
  - Likert scale: Numerical but normalized

- Category (Nominal)
  - No number from single measurement
  - Frequency data used from whole data set

Usability Questionnaire (almost standard)*

- **Usefulness**
  - It helps me be more effective.
  - It helps me be more productive.
  - It is useful.
  - It gives me more control over the activities in my life.
  - It makes the things I want to accomplish easier to get done.
  - It saves me time when I use it.
  - *It meets my needs.*
  - It does everything I would expect it to do.

- **Ease of Use**
  - It is easy to use.
  - It is simple to use.
  - It is user friendly.
  - It requires the fewest steps possible to accomplish what I want to do with it.
  - *It is flexible.*
  - *Using it is effortless.*
  - *I can use it without written instructions.*
  - *I don't notice any inconsistencies as I use it.*
  - *Both occasional and regular users would like it.*
  - *I can recover from mistakes quickly and easily.*
  - *I can use it successfully every time.*

Usability Questionnaire (almost standard)*

- **Ease of Learning**
  - I learned to use it quickly.
  - I easily remember how to use it.
  - It is easy to learn to use it.
  - *I quickly became skillful with it.*

- **Satisfaction**
  - I am satisfied with it.
  - I would recommend it to a friend.
  - It is fun to use.
  - It works the way I want it to work.
  - It is wonderful.
  - I feel I need to have it.
  - It is pleasant to use.

Analysis type

- Parametric
  - Based on assumption that data you have collected is normally distributed (slight deviation is ok)
  - For interval scale
  - t-tests for comparing two means
  - ANOVA for factorial design (F-test)
  - Unrelated t-test (matched pair)
  - Pearson (correlation)

- Non parametric
  - Ordinal / Likert
  - Categorical
  - Wilcoxon signed rank test
  - Shearman Rho (correlation)
  - Chi square test (matched pair)

Statistical Test

1. Null hypothesis: a contradiction of the alternative hypothesis

2. Alternative hypothesis: the hypothesis the researcher wants to support.

3. Test statistic and its *p*-value: sample evidence calculated from sample data.

4. Rejection region—critical values and significance levels: values that separate rejection and nonrejection of the null hypothesis

5. Conclusion: Reject or do not reject the null hypothesis, stating the practical significance of your conclusion.

Errors and Statistical Significance

1. The significance level α is the probability if rejecting H_0 when it is in fact true.

2. The *p*-value is the probability of observing a test statistic as extreme as or more than the one observed; also, the smallest value of α for which H_0 can be rejected.

3. When the *p*-value is less than the significance level α, the null hypothesis is rejected. This happens when the test statistic exceeds the critical value.

Example

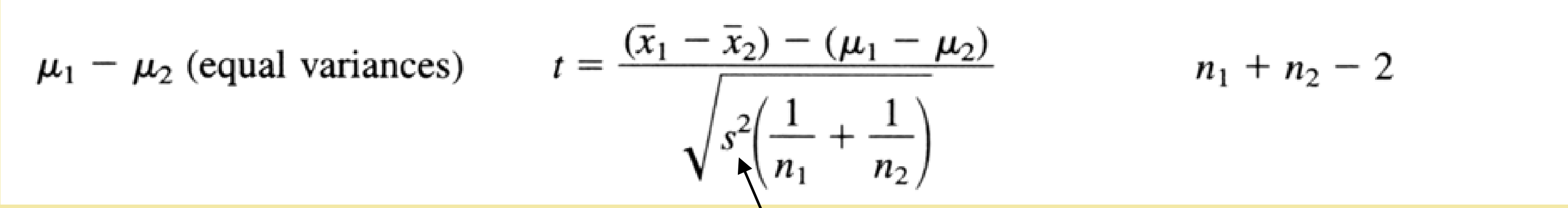
Steve and Mary both developed a UI for a Chinese fast food restaurant management software. Two groups of people tried out the UI for a particular task, and the length of time to complete the task is shown in the table below.

Do the data present sufficient evidence to indicate that the mean time to finish the task is actually different?

UI # 1	**UI #2**
32	35
37	31
35	29
28	25
41	34
44	40
35	27
31	32
34	31

t-test

- Independent variable (Factor): Which UI?
- Dependent variable: Task completion time
- Number of samples: Lower than 30
  - Student t test
  - When you have large number of samples: Z test

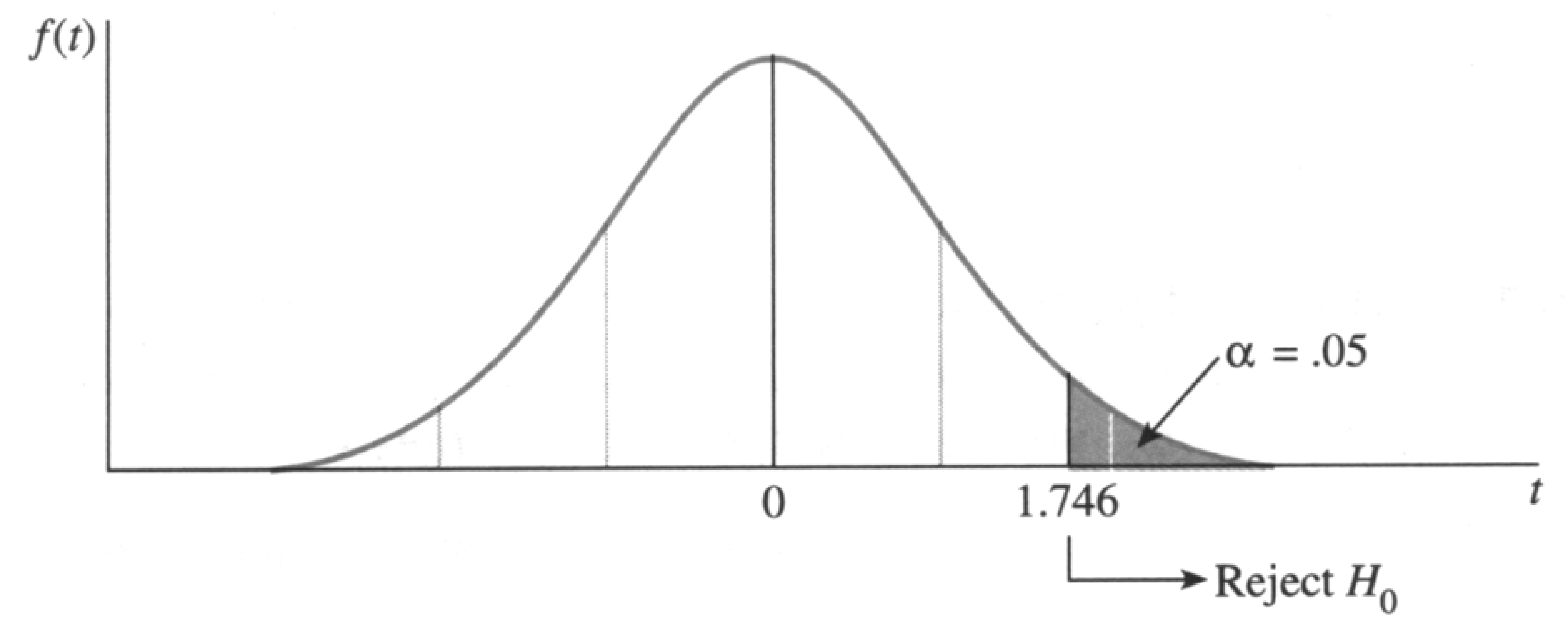


$\mu_1 - \mu_2$ (equal variances)	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$	$n_1 + n_2 - 2$

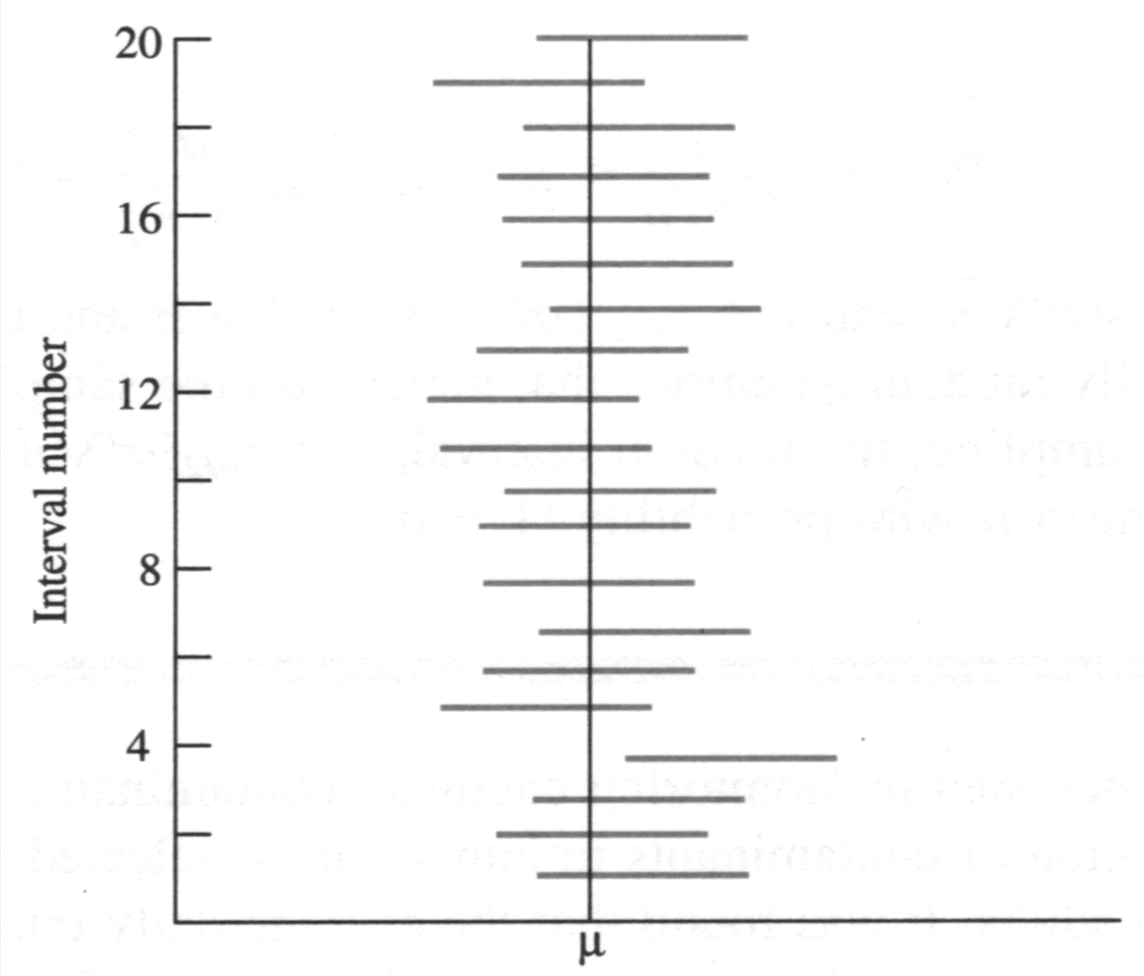
t = 1.76

$p > 0.05$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$


$f(t)$
$\alpha = .05$
0
1.746
t
Reject H_0

Concept of Confidence Level of the Experiment

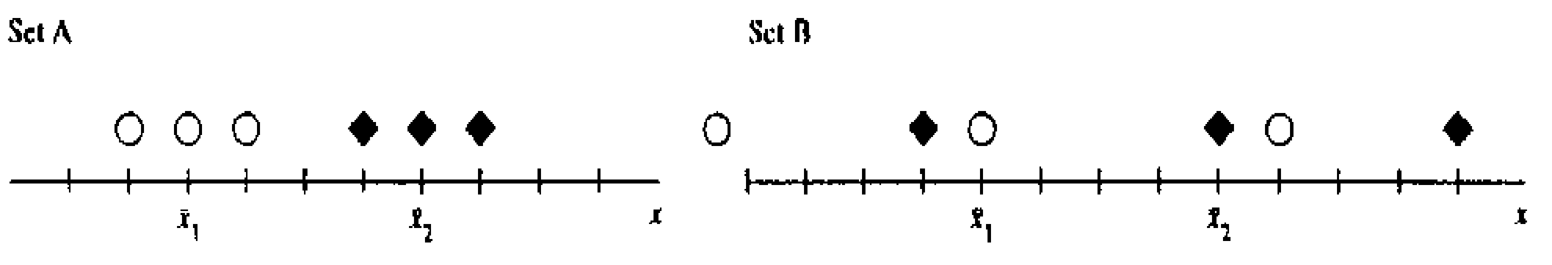


What Is an Analysis of Variance?

- Responses exhibit variability.

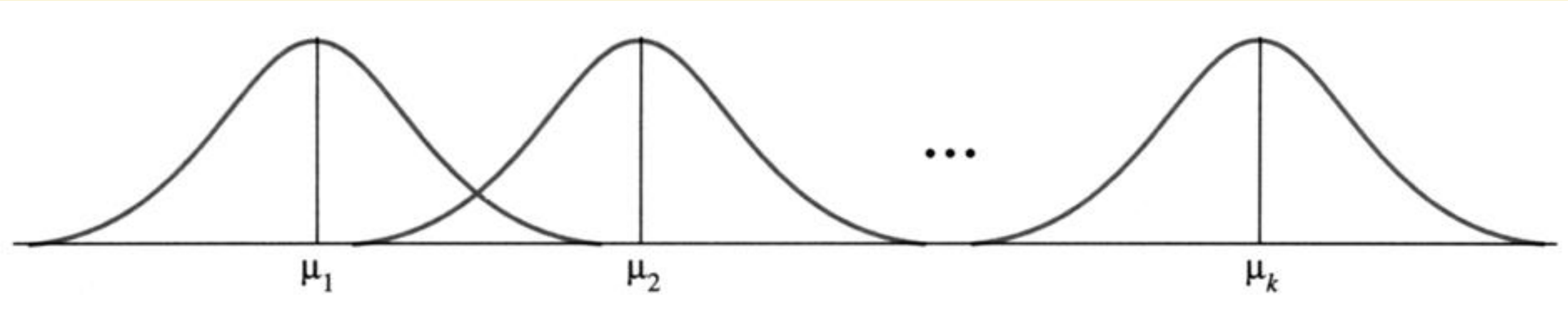
- In an analysis of variance (ANOVA), the total variation in the response measurements is divided into portions that may be attributed to various factors

- The variability of the measurements within experimental groups and between experimental groups and their comparisons is formalized by the ANOVA.

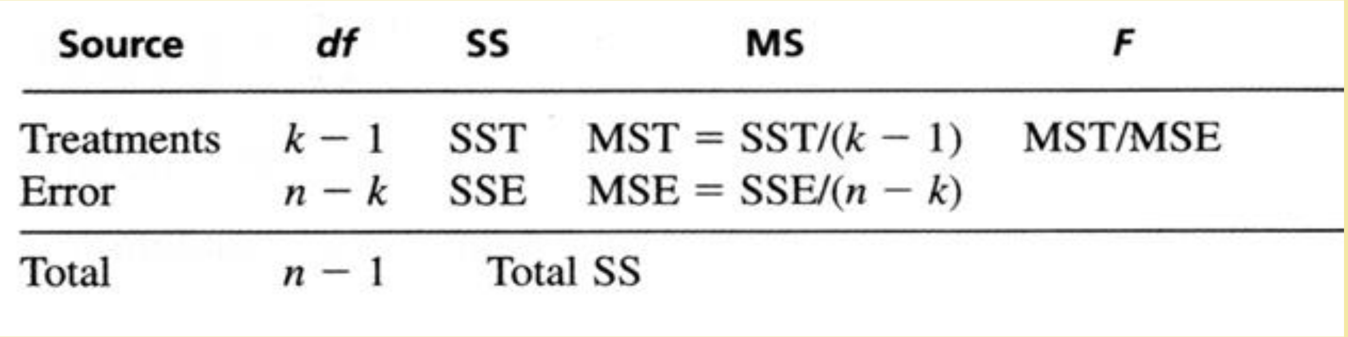


Analysis of Variance for a Completely Randomized Design

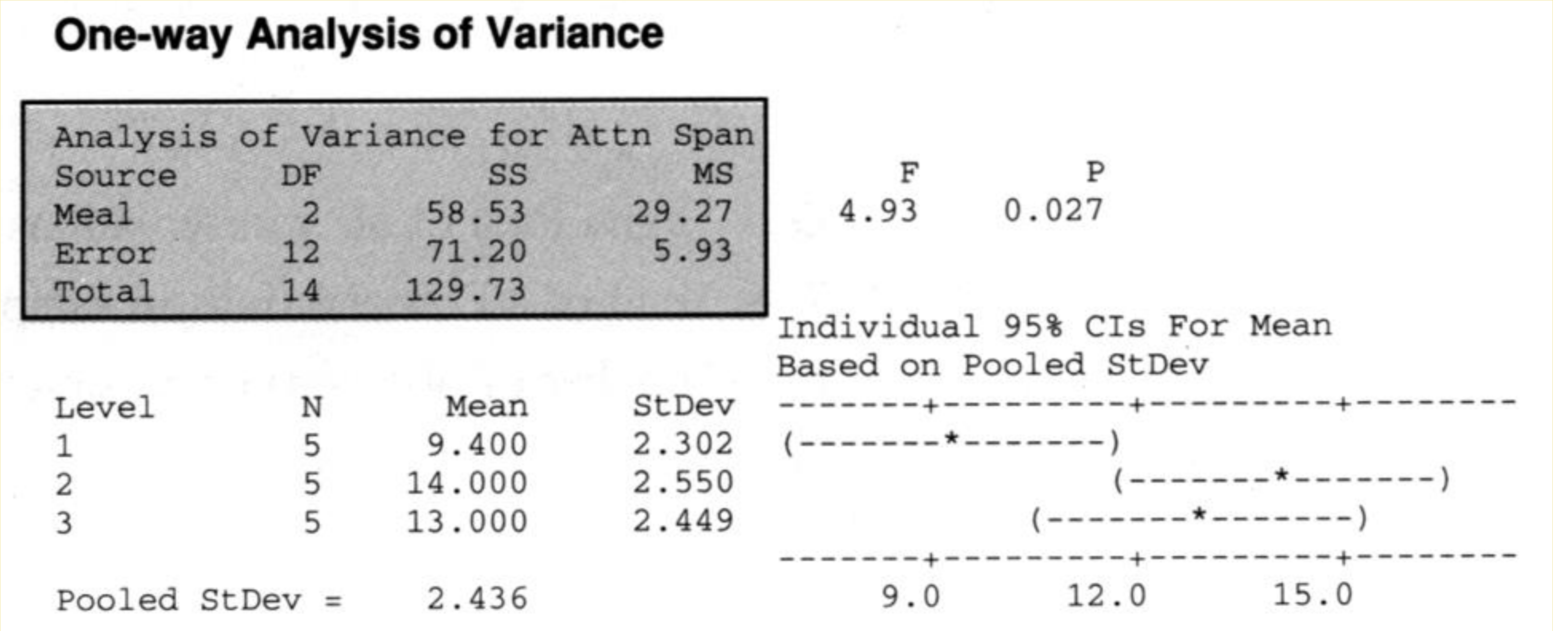
- Suppose you want to compare *k* population means based on independent random samples of size n_1, n_2, …, n_K from normal populations with a common variance σ^2. They have the same shape, but potentially different locations:



ANOVA Table for *k* Independent Random Samples: Completely Randomized Design



Source	*df*	SS	MS	*F*
Treatments	$k - 1$	SST	MST = SST/$(k - 1)$	MST/MSE
Error	$n - k$	SSE	MSE = SSE/$(n - k)$	
Total	$n - 1$	Total SS		



One-way Analysis of Variance

```
Analysis of Variance for Attn Span
Source      DF        SS        MS          F          P
Meal         2     58.53     29.27       4.93      0.027
Error       12     71.20      5.93
Total       14    129.73
                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
Level        N      Mean     StDev  -------+---------+---------+--------
1            5     9.400     2.302  (-------*-------)
2            5    14.000     2.550                 (-------*-------)
3            5    13.000     2.449             (-------*-------)
                                    -------+---------+---------+--------
Pooled StDev =     2.436                  9.0      12.0      15.0
```

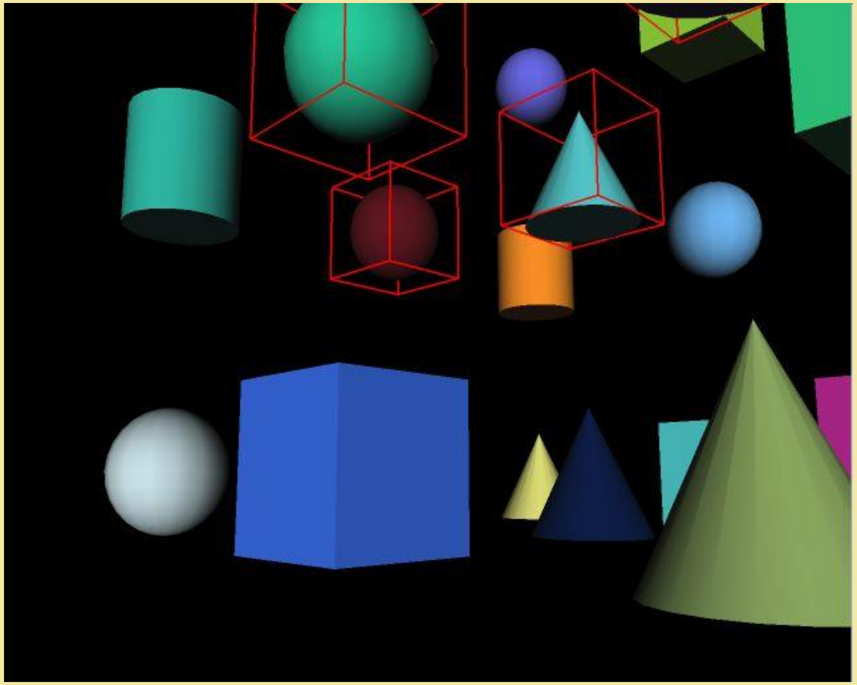
Experiment Setups

Motion-based hand-held VR

Button-based hand-held VR

Mouse/Keyboard interaction in small screen/desktop/large screen

Tasks in the Experiments



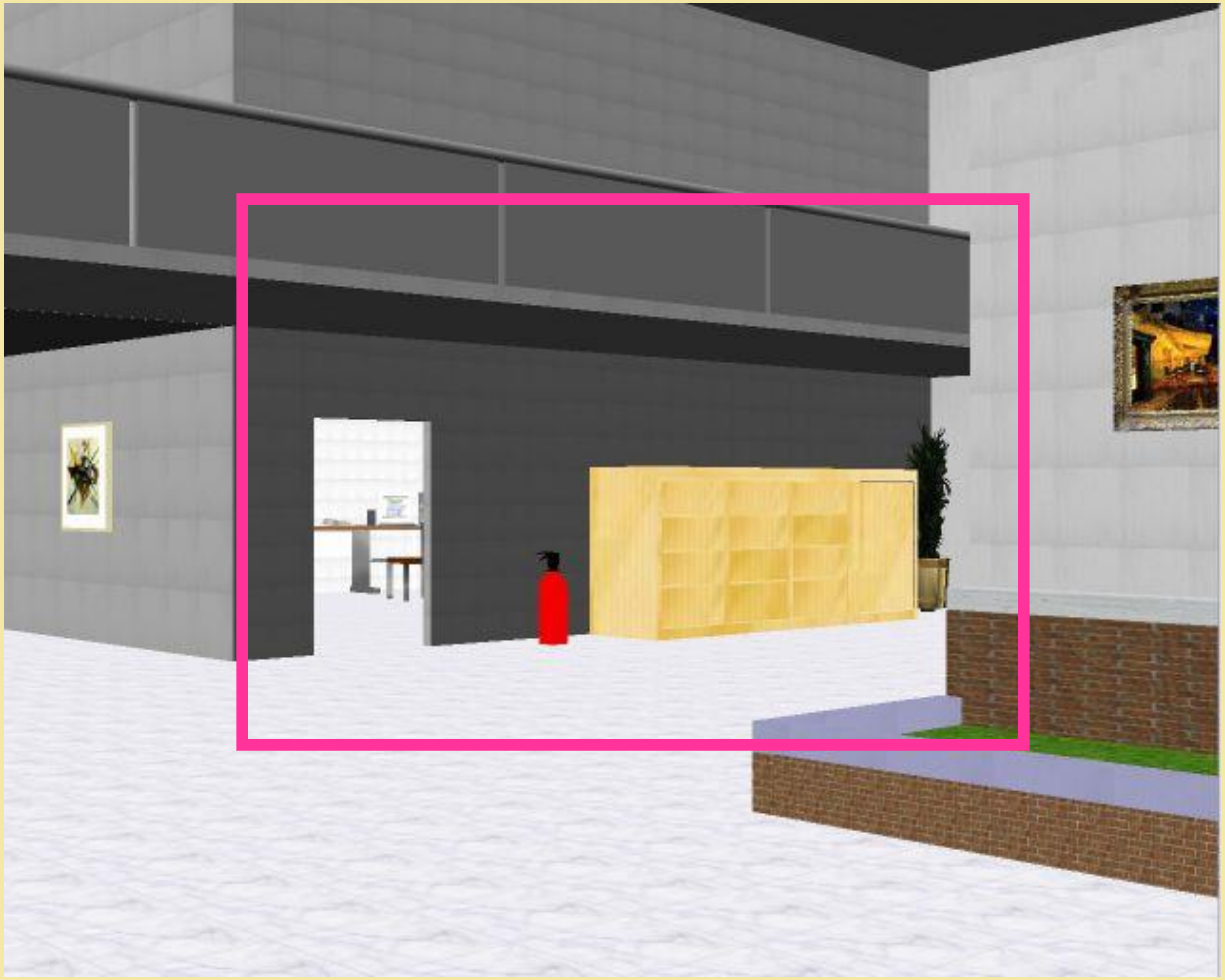
Navigating in virtual environment
➔Usability, presence/immersion, enjoyment, and perceived FOV

Locating and selecting objects
➔ Task performance

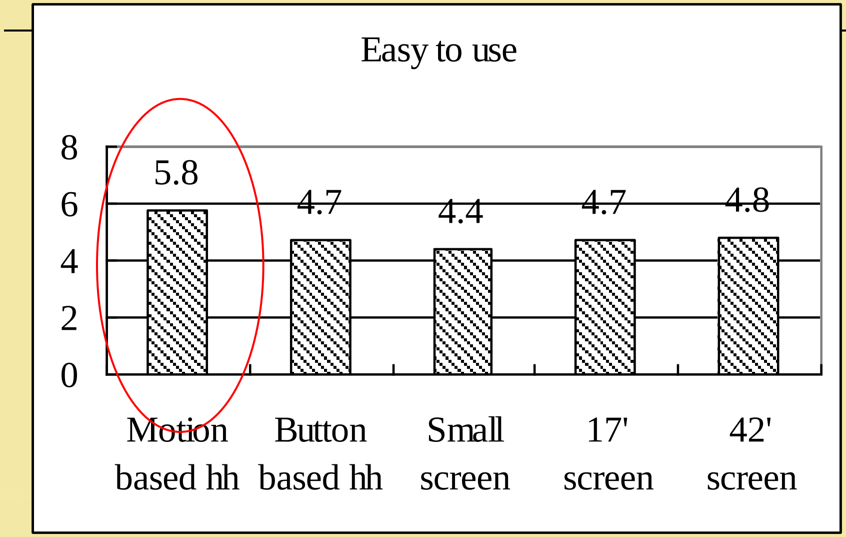
Dependent Variables

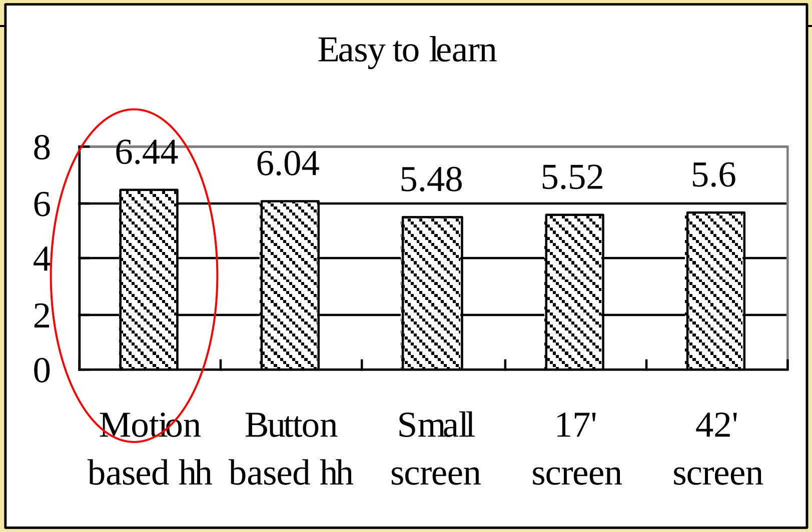
- Questionnaire answers
  - Visual quality and depth perception
  - Auditory quality
  - Usability
  - Presence (modified SUS) / Immersion
  - Distraction
  - Enjoyment
  - Cyber-sickness
  - FOV
- Task performance
- Perceived FOV

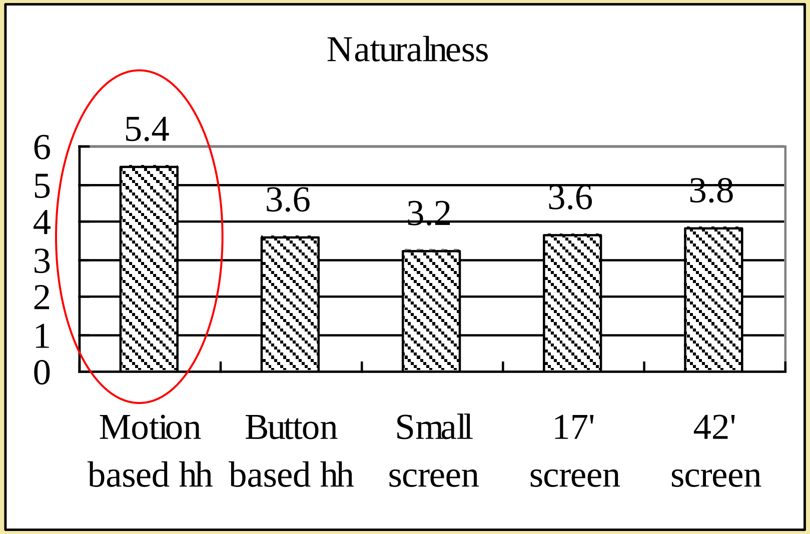
Measuring perceived FOV

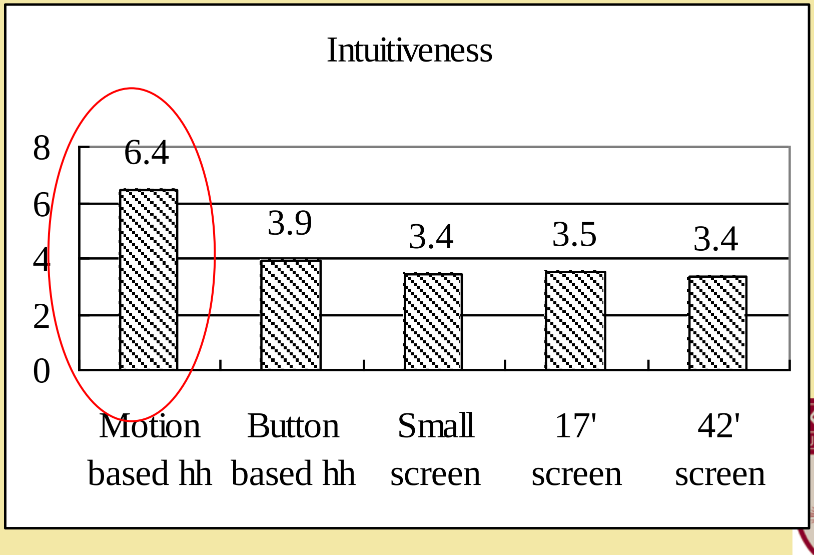


Usability

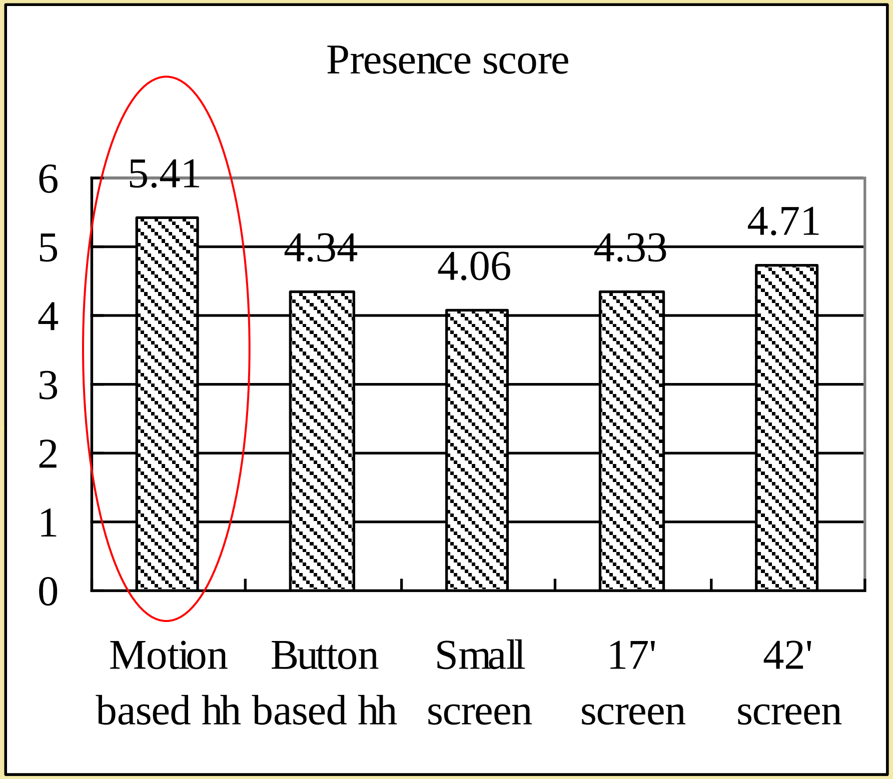

Easy to use

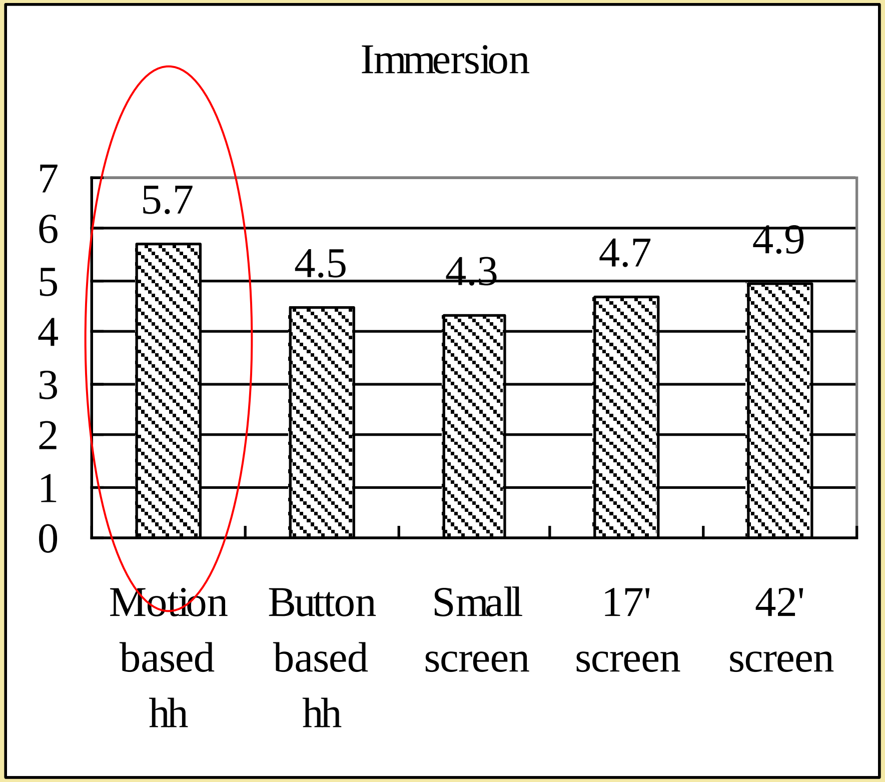

Easy to learn


Naturalness

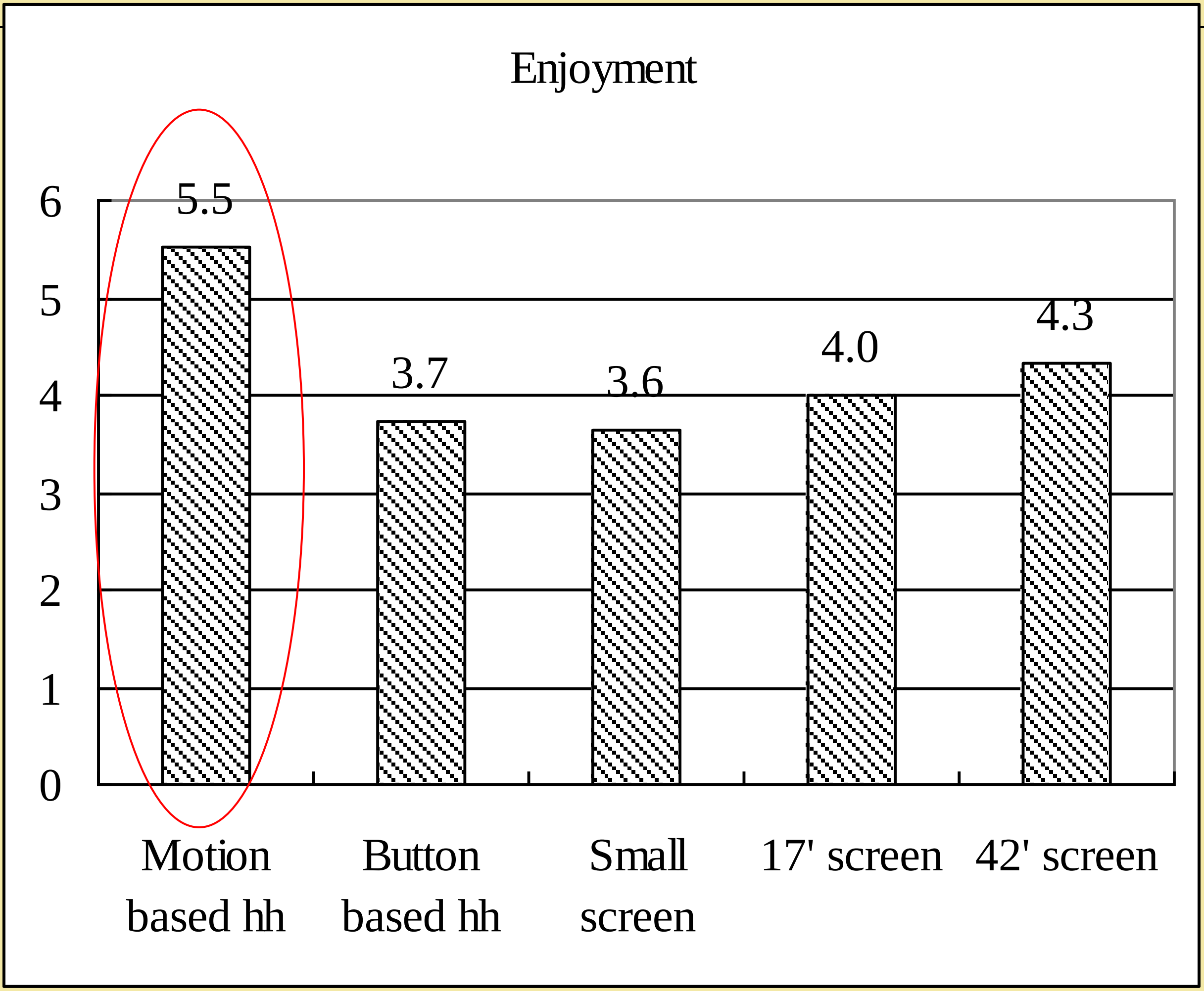

Intuitiveness

Presence/Immersion

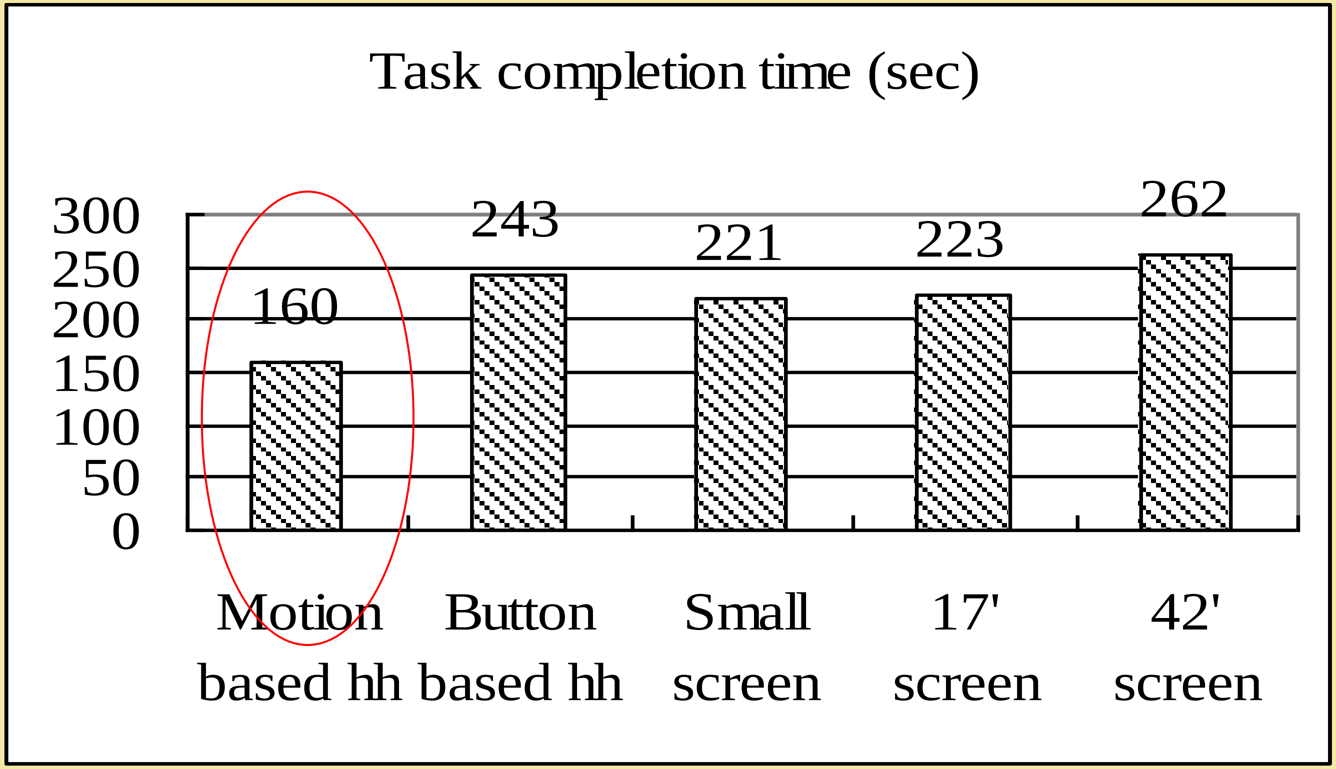




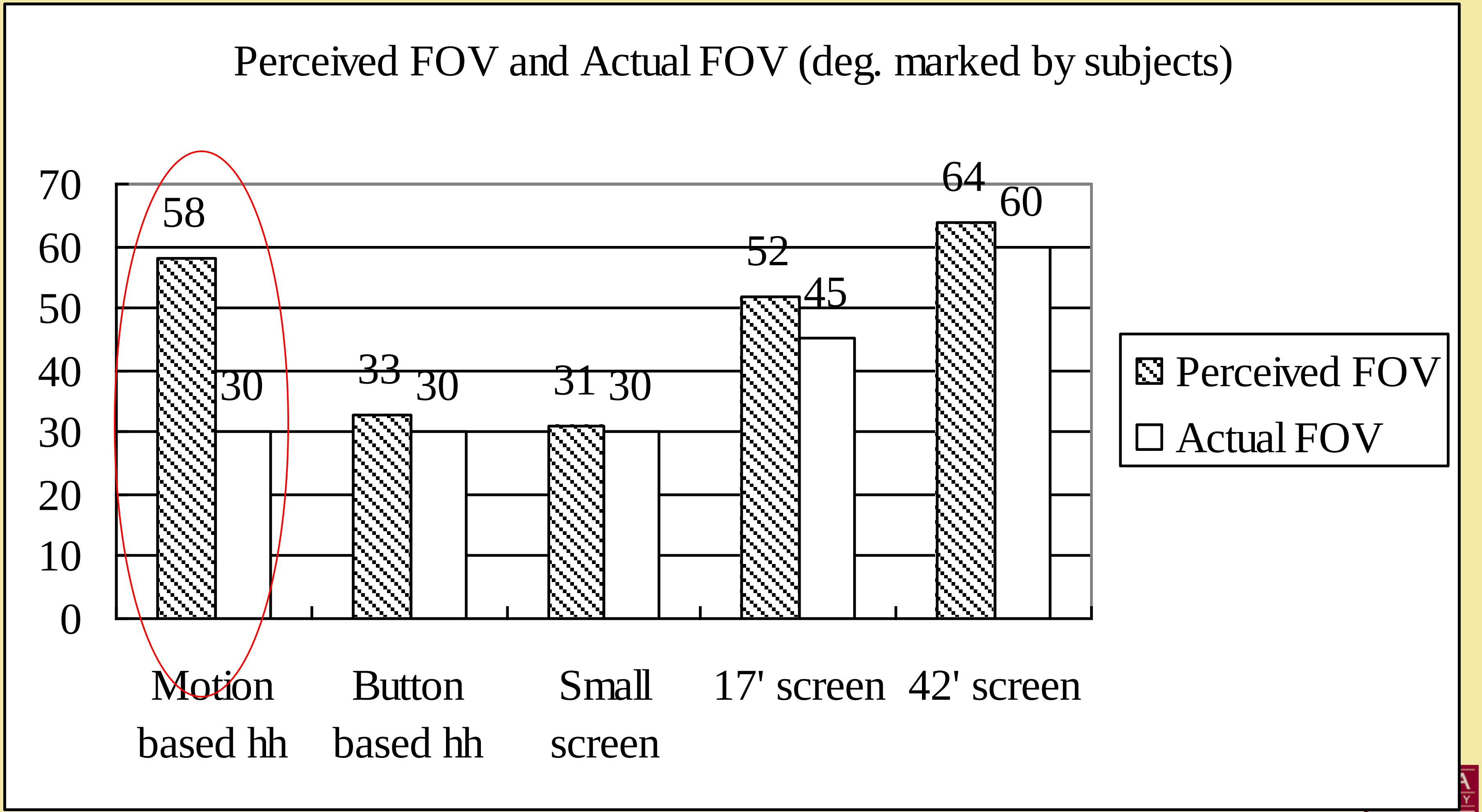
Enjoyment

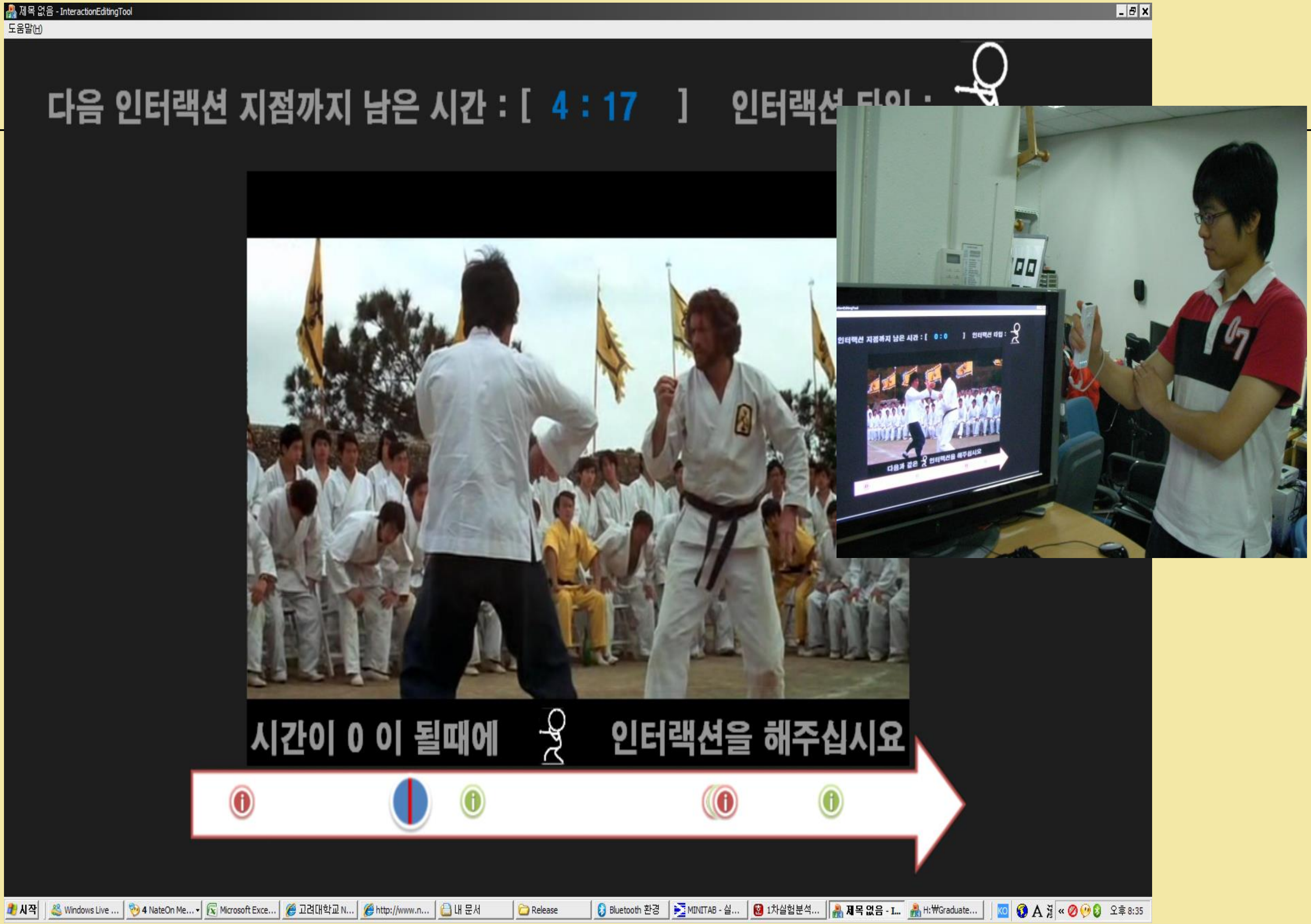

Enjoyment
6
5
4
3
2
1
0
5.5
3.7
3.6
4.0
4.3
Motion based hh
Button based hh
Small screen
17' screen
42' screen

Task Performance



Perceived FOV

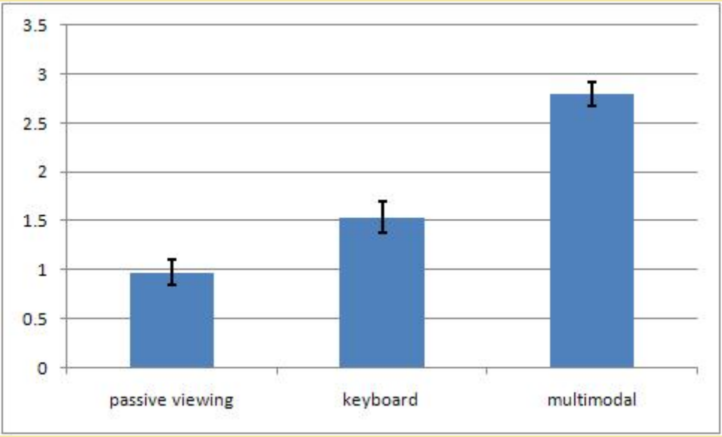


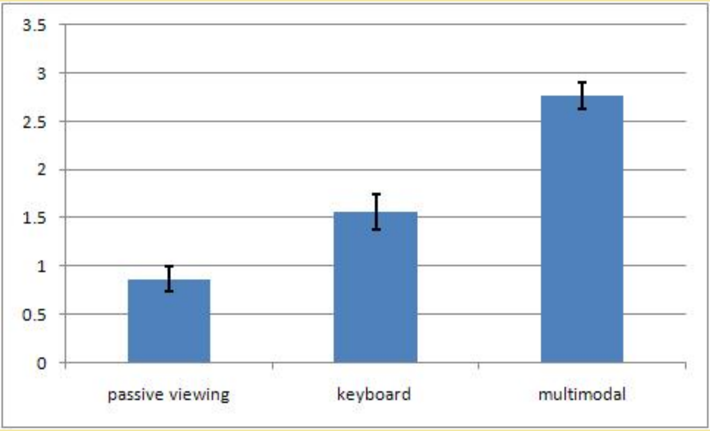


제목 없음 - InteractionEditingTool

다음 인터랙션 지점까지 남은 시간 : [4 : 17] 인터랙션 타입 :

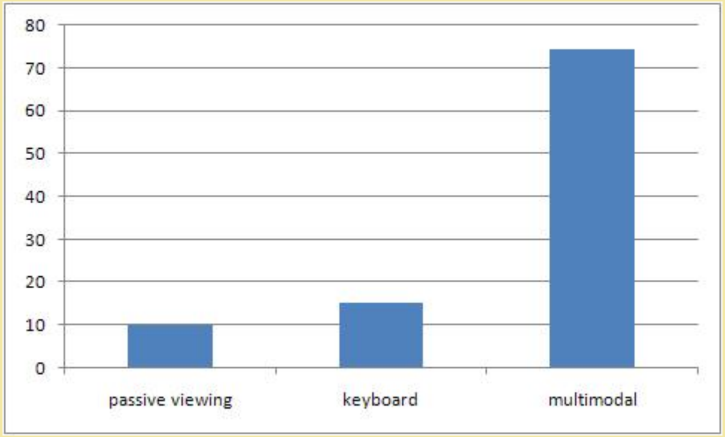
시간이 0 이 될때에 인터랙션을 해주십시요

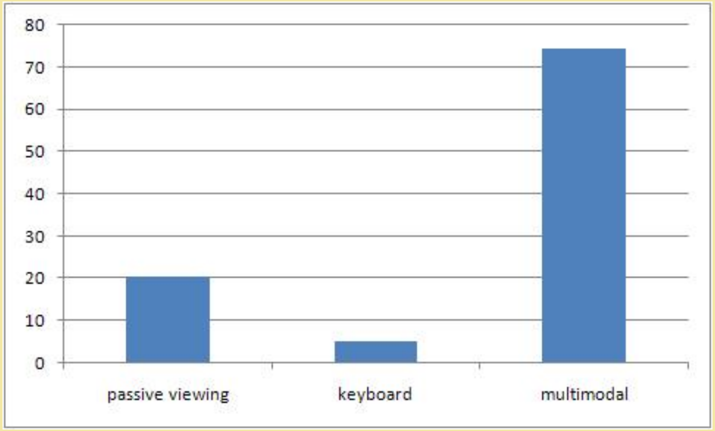




User Empathy

Presence





General Preference

Preference/Immersion

$p < 0.0000$

Acceptance Test

- For large implementation projects, the customer or manager usually sets objective and measurable goals for hardware and software performance.

- If the completed product fails to meet these acceptance criteria, the system must be reworked until success is demonstrated.

- Rather than the vague and misleading criterion of "user friendly," measurable criteria for the user interface can be established for the following:
  - Time to learn specific functions
  - Speed of task performance
  - Rate of errors by users
  - Human retention of commands over time
  - Subjective user satisfaction