# *EVALUASI EMPIRIS*

Pengenalan Evaluasi Empiris
Perancangan Eksperimen
Partisipasi, IRB dan Etika
Pengumpulan Data
Analisa Data dan Interpretasi Hasil
Penggunaan Hasil Rancangan

# Why Evaluate?

Recall:

- Users and their tasks were identified
- Needs and requirements were specified
- Interface was designed, prototype built
- *But is it any good? Does the system support the users in their tasks? Is it better than what was there before (if anything)?*

# Types of Evaluation

- Interpretive and Predictive (a reminder)
  - Heuristic evaluation, cognitive walkthroughs, ethnography…
- Summative vs. Formative
  - What were they, again?

# Now With Users Involved

- Interpretive (naturalistic) vs. Empirical:

- Naturalistic
  - In realistic setting, usually includes some detached observation, careful study of users
- Empirical
  - People use system, manipulate independent variables and observe dependent ones

# Why Gather Data?

- Design the experiment to collect the data to test the hypotheses to evaluate the interface to refine the design

- Information gathered can be:
  *objective* or *subjective*
- Information also can be:
  *qualitative* or *quantitative*

# Conducting an Experiment

- Determine the TASK
- Determine the performance measures
- Develop the experiment
- IRB approval
- Recruit participants
- Collect the data
- Inspect & analyze the data
- Draw conclusions to resolve design problems
- Redesign and implement the revised interface

# The Task

- Benchmark tasks - gather quantitative data
- Representative tasks - add breadth, can help understand process
- Tell them what to do, not how to do it
- Issues:
  - Lab testing vs. field testing
  - Validity - typical users; typical tasks; typical setting?
  - Run pilot versions to shake out the bugs

# "Benchmark" Tasks

- Specific, clearly stated task for users to carry out
- Example:  Email handler
  - "Find the message from Mary and reply with a response of 'Tuesday morning at 11'."
- Users perform these under a variety of conditions and you measure performance

# Defining Performance

- Based on the task
- Specific, objective measures/metrics
- Examples:
  - Speed (reaction time, time to complete)
  - Accuracy (errors, hits/misses)
  - Production (number of files processed)
  - Score (number of points earned)
  - …others…?

# Types of Variables

- Independent
  - What you're studying, what you intentionally vary (e.g., interface feature, interaction device, selection technique)
- Dependent
  - Performance measures you record or examine (e.g., time, number of errors)

# “Controlling” Variables

- Prevent a variable from affecting the results in any systematic way
- Methods of controlling for a variable:
  - Don’t allow it to vary
    - e.g., all males
  - Allow it to vary randomly
    - e.g., randomly assign participants to different groups
  - Counterbalance - systematically vary it
    - e.g., equal number of males, females in each group
  - The appropriate option depends on circumstances

# Hypotheses

- What you predict will happen
- More specifically, the way you predict the dependent variable (i.e., accuracy) will depend on the independent variable(s)
- "Null" hypothesis ($H_o$)
  - Stating that there will be no effect
  - e.g., "There will be no difference in performance between the two groups"
  - Data used to try to disprove this null hypothesis

# Example

- Do people complete operations faster with a black-and-white display or a color one?
  - Independent - display type (color or b/w)
  - Dependent - time to complete task (minutes)
  - Controlled variables - same number of males and females in each group
  - Hypothesis: Time to complete the task will be shorter for users with color display
  - $H_o$: $Time_{color} = Time_{b/w}$
  - Note: Within/between design issues, next

# Experimental Designs

- Within Subjects Design
  - Every participant provides a score for all levels or conditions

| | Color | B/W |
|---|---|---|
| P1 | 12 secs. | 17 secs. |
| P2 | 19 secs. | 15 secs. |
| P3 | 13 secs. | 21 secs. |
| ... | | |

- Between Subjects
  - Each participant provides results for only one condition

| Color | | B/W | |
|---|---|---|---|
| P1 | 12 secs. | P2 | 17 secs. |
| P3 | 19 secs. | P5 | 15 secs. |
| P4 | 13 secs. | P6 | 21 secs. |
| ... | | | |

# Within Subjects Designs

- More efficient:
  - Each subject gives you more data - they complete more “blocks” or “sessions”
- More statistical “power”:
  - Each person is their own control
- Therefore, can require fewer participants
- May mean more complicated design to avoid “order effects”
  - e.g. seeing color then b/w may be different from seeing b/w then color

# Between Subjects Designs

- Fewer order effects
  - Participant may learn from first condition
  - Fatigue may make second performance worse
- Simpler design & analysis
- Easier to recruit participants (only one session)
- Less efficient

# IRB, Participants, & Ethics

- Institutional Review Board (IRB)
  - http://www.osp.gatech.edu/compliance.htm
- Reviews all research involving human (or animal) participants
- Safeguarding the participants, and thereby the researcher and university
- Not a science review (i.e., not to asess your research ideas); only safety & ethics
- Complete Web-based forms, submit research summary, sample consent forms, etc.
- All experimenters must complete NIH online history/ethics course prior to submitting

# Recruiting Participants

- Various “subject pools”
  - Volunteers
  - Paid participants
  - Students (e.g., psych undergrads) for course credit
  - Friends, acquaintances, family, lab members
  - “Public space” participants - e.g., observing people walking through a museum
- Must fit user population (validity)
- Motivation is a big factor - not only $$ but also explaining the importance of the research
- Note: Ethics, IRB, Consent apply to *all* participants, including friends & “pilot subjects”

# Ethics

- Testing can be arduous
- Each participant should consent to be in experiment (informal or formal)
  - Know what experiment involves, what to expect, what the potential risks are
- Must be able to stop without danger or penalty
- All participants to be treated with respect

# Consent

- Why important?
  - People can be sensitive about this process and issues
  - Errors will likely be made, participant may feel inadequate
  - May be mentally or physically strenuous
- What are the potential risks (there are always risks)?
  - Examples?
- “Vulnerable” populations need special care & consideration (& IRB review)
  - Children; disabled; pregnant; students (why?)

# Attribution Theory

- Studies why people believe that they succeeded or failed--themselves or outside factors (gender, age differences)

- Explain how errors or failures are not participant's problem---places where interface needs to be improved

# Evaluation is Detective Work

- Goal: gather evidence that can help you determine whether your hypotheses are correct or not.
- Evidence (data) should be:
  - Relevant
  - Diagnostic
  - Credible
  - Corroborated

# Data as Evidence

- Relevant
  - Appropriate to address the hypotheses
    - e.g., Does measuring “number of errors” provide insight into how effective your new air traffic control system supports the users’ tasks?
- Diagnostic
  - Data unambiguously provide evidence one way or the other
    - e.g., Does asking the users’ preferences clearly tell you if the system performs better? (Maybe)

# Data as Evidence

- Credible
  - Are the data trustworthy?
    - Gather data carefully; gather enough data
- Corroborated
  - Do more than one source of evidence support the hypotheses?
    - e.g., Both accuracy and user opinions indicate that the new system is better than the previous system. But what if completion time is slower?

# General Recommendations

- Include both objective & subjective data
  - e.g., “completion time” and “preference”
- Use multiple measures, within a type
  - e.g., “reaction time” and “accuracy”
- Use quantitative measures where possible
  - e.g., preference score (on a scale of 1-7)

Note: Only gather the data required; do so with the min. interruption, hassle, time, etc.

# Types of Data to Collect

- "Demographics"
  - Info about the participant, used for grouping or for correlation with other measures
    - e.g., handedness; age; first/best language; SAT score
    - Note: Gather if it is relevant. Does not have to be self-reported: you can use tests (e.g.,Edinburgh Handedness)
- Quantitative data
  - What you measure
    - e.g., reaction time; number of yawns
- Qualitative data
  - Descriptions, observations that are not quantified
    - e.g., different ways of holding the mouse; approaches to solving problem; trouble understanding the instructions

# Planning for Data Collection

- What data to gather?
  - Depends on the task and any benchmarks
- How to gather the data?
  - Interpretive, natural, empirical, predictive??
- What criteria are important?
  - Success on the task? Score? Satisfaction?…
- What resources are available?
  - Participants, prototype, evaluators, facilities, team knowledge (programming, stats, etc.)

# Collecting Data

- Capturing the Session
  - Observation & Note-taking
  - Audio and video recording
  - Instrumented user interface
  - Software logs
  - Think-aloud protocol - can be very helpful
  - Critical incident logging - positive & negative
- Post-session activities
  - Structured interviews; debriefing
    - “What did you like best/least?”; “How would you change..?”
  - Questionnaires, comments, and rating scales
  - Post-hoc video coding/rating by experimenter

# Observing Users

- Not as easy as you think
- One of the best ways to gather feedback about your interface
- Watch, listen and learn as a person interacts with your system

# Observation

- Direct
  - In same room
  - Can be intrusive
  - Users aware of your presence
  - Only see it one time
  - May use 1-way mirror to reduce intrusion
  - Cheap, quicker to set up and to analyze
- Indirect
  - Video recording
  - Reduces intrusion, but doesn't eliminate it
  - Cameras focused on screen, face & keyboard
  - Gives archival record, but can spend a lot of time reviewing it

# Location

- Observations may be
  - In lab - Maybe a specially built usability lab
    - Easier to control
    - Can have user complete set of tasks
  - In field
    - Watch their everyday actions
    - More realistic
    - Harder to control other factors

# Challenge

- In simple observation, you observe actions but don’t know what’s going on in their head

- Often utilize some form of *verbal protocol* where users describe their thoughts

# Verbal Protocol

- One technique: *Think-aloud*
  - User describes verbally what s/he is thinking while performing the tasks
    - What they believe is happening
    - Why they take an action
    - What they are trying to do
- Very widely used, useful technique
- Allows you to understand user's thought processes better
- Potential problems:
  - Can be awkward for participant
  - Thinking aloud can modify way user performs task

# Teams

- Another technique: *Co-discovery learning* (Constructive interaction)
  - Join pairs of participants to work together
  - Use think aloud
  - Perhaps have one person be semi-expert (coach) and one be novice
  - More natural (like conversation) so removes some awkwardness of individual think aloud

# Alternative

- What if thinking aloud during session will be too disruptive?
- Can use *post-event protocol*
  - User performs session, then watches video and describes what s/he was thinking
  - Sometimes difficult to recall
  - Opens up door of interpretation

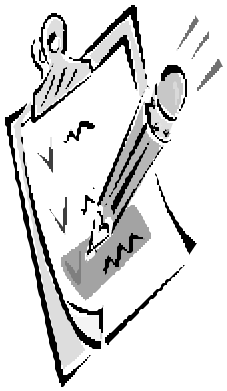
# Historical Record

- In observing users, how do you capture events in the session for later analysis?

# Capturing a Session

## 1. Paper & pencil

- Can be slow
- May miss things
- Is definitely cheap and easy



| | Task 1 | Task 2 | Task 3 | ... |
|---|---|---|---|---|
| Time 10:00 | | S | | |
| 10:03 | | e | | |
| 10:08 | | | S | |
| 10:22 | | | e | |

# Capturing a Session

2. Recording (audio and/or video)
   - Good for talk-aloud
   - Hard to tie to interface
   - Multiple cameras probably needed
   - Good, rich record of session
   - Can be intrusive
   - Can be painful to transcribe and analyze

# Capturing a Session

3. Software logging
   - Modify software to log user actions
   - Can give time-stamped keypress or mouse event
   - Two problems:
     - Too low-level, want higher level events
     - Massive amount of data, need analysis tools

# Subjective Data

- Satisfaction is an important factor in performance over time
- Learning what people prefer is valuable data to gather

# Methods

- Ways of gathering subjective data
  - Questionnaires
  - Interviews
  - Booths (e.g., trade show)
  - Call-in product hot-line
  - Field support workers
- (Focus on first two)

# Questionnaires

- Preparation is expensive, but administration is cheap
- Oral vs. written
  - Oral advs: Can ask follow-up questions
  - Oral disadvs: Costly, time-consuming
- Forms can provide more quantitative data
- Issues
  - Only as good as questions you ask
  - Establish purpose of questionnaire
  - Don’t ask things that you will not use
  - Who is your audience?
  - How do you deliver and collect questionnaire?

# Questionnaire Topic

- Can gather demographic data and data about the interface being studied
- Demographic data:
  - Age, gender
  - Task expertise
  - Motivation
  - Frequency of use
  - Education/literacy

# Interface Data

- Can gather data about
  - screen
  - graphic design
  - terminology
  - capabilities
  - learning
  - overall impression
  - ...

# Closed Format

- Closed format
  - Answer restricted to a set of choices
  - Typically very quantifiable
  - Variety of styles
- Likert Scale
  - Typical scale uses 5, 7 or 9 choices
  - Above that is hard to discern
  - Doing an odd number gives the neutral choice in the middle
  - You may not want to give a neutral option

Characters on screen were:

hard to read easy to read

1 2 3 4 5 6 7

# Other Styles

Which word processing systems do you use?

- ☐ LaTeX
- ☐ Word
- ☐ FrameMaker
- ☐ WordPerfect

Rank from
1 - Very helpful
2 - Ambivalent
3 - Not helpful
0 - Unused

___ Tutorial
___ On-line help
___ Documentation

# Open Format

- Asks for unprompted opinions
- Good for general, subjective information, but difficult to analyze rigorously
- May help with design ideas
  - "Can you suggest improvements to this interface?"

# Closed Format

- Advantages
  - Clarify alternatives
  - Easily quantifiable
  - Eliminate useless answer
- Disadvantages
  - Must cover whole range
  - All should be equally likely
  - Don't get interesting, "different" reactions

# Questionnaire Issues

- Question specificity
  - "Do you have a computer?"
- Language
  - Beware terminology, jargon
- Clarity
  - "How effective was the system?" (ambiguous)
- Leading questions
  - Can be phrased either positive or negative
- Prestige bias - (British sex survey)
  - People answer a certain way because they want you to think that way about them
- Embarrassing questions
  - "What did you have the most problem with?"
- Hypothetical questions
- "Halo effect
  - When estimate of one feature affects estimate of another (eg, intelligence/looks)
  - Aesthetics & usability, one example in HCI

# Deployment

- Steps
  - Discuss questions among team
  - Administer verbally/written to a few people (pilot). Verbally query about thoughts on questions
  - Administer final test
  - Use computer-based input if possible
  - Have data pre-processed, sorted, set up for later analysis at the time it is collected

# Interviews

- Get user's viewpoint directly, but certainly a subjective view

- Advantages:
  - Can vary level of detail as issue arises
  - Good for more exploratory type questions which may lead to helpful, constructive suggestions

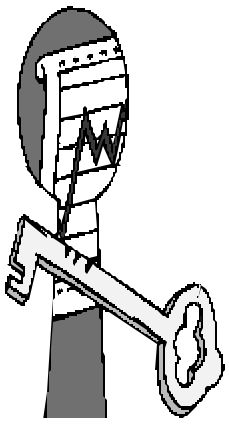
- Disadvantages
  - Subjective view
  - Interviewer(s) can bias the interview
  - Problem of inter-rater or inter-experimenter reliability (a stats term meaning agreement)
  - User may not appropriately characterize usage
  - Time-consuming
  - Hard to quantify

# Interview Process

- How to be effective
  - Plan a set of questions (provides for some consistency)
  - Don’t ask leading questions
    - “Did you think the use of an icon there was really good?”
- Can be done in groups
  - Get consensus, get lively discussion going

# Data Inspection

- Look at the results
- First look at each participant's data
  - Were there outliers, people who fell asleep, anyone who tried to mess up the study, etc.?
- Then look at aggregate results and descriptive statistics



# Inspecting Your Data

- “What happened in this study?”
- Keep in mind the goals and hypotheses you had at the beginning
- Questions:
  - Overall, how did people do?
  - “5 W’s” (Where, what, why, when, and for whom were the problems?)

# Descriptive Statistics

- For all variables, get a feel for results:
- Total scores, times, ratings, etc.
- Minimum, maximum
- Mean, median, ranges, etc.

❖ e.g. “Twenty participants completed both sessions (10 males, 10 females; mean age 22.4, range 18-37 years).”

❖ e.g. “The median time to complete the task in the mouse-input group was 34.5 s (min=19.2, max=305 s).”

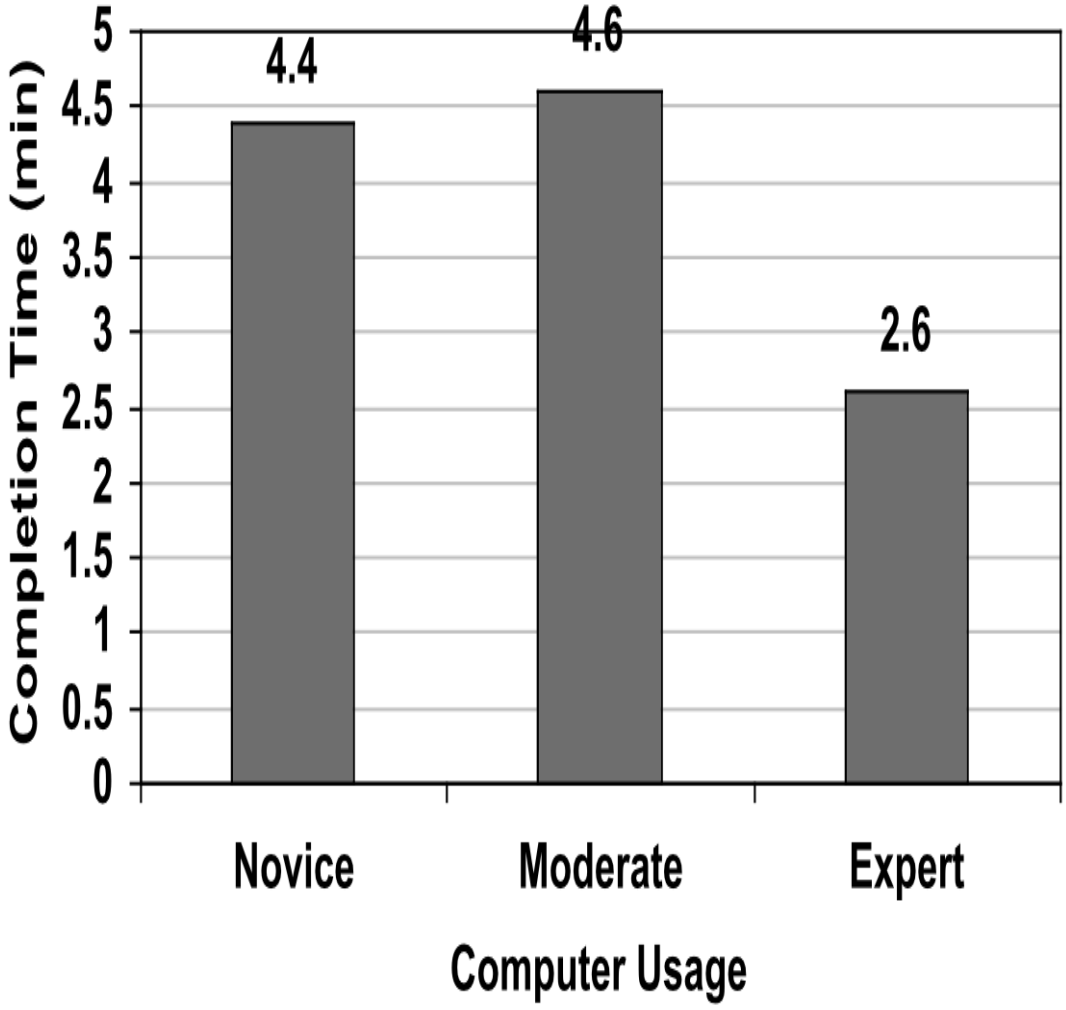
What is the difference between mean & median? Why use one or the other?

# Subgroup Stats

- Look at descriptive stats (means, medians, ranges, etc.) for any subgroups
  - e.g. “The mean error rate for the mouse-input group was 3.4%. The mean error rate for the keyboard group was 5.6%.”
  - e.g. “The median completion time (in seconds) for the three groups were: novices: 4.4, moderate users: 4.6, and experts: 2.6.”

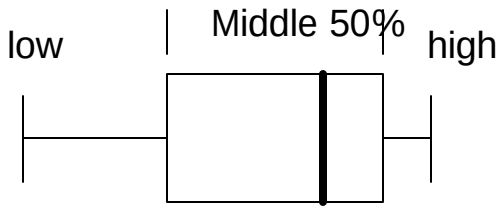
# Plot the Data

- Look for the trends graphically



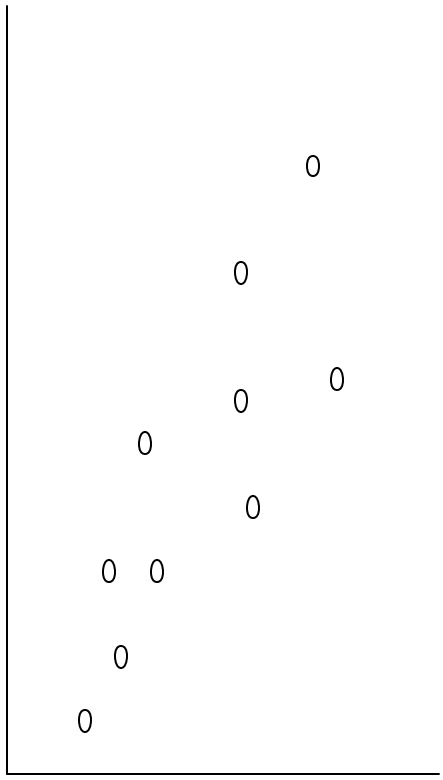
# Other Presentation Methods

Box plot





Scatter plot



# Experimental Results

- How does one know if an experiment's results mean anything or confirm any beliefs?

- Example: 40 people participated,
  28 preferred interface 1,
  12 preferred interface 2
- What do you conclude?

# Inferential (Diagnostic) Stats

- Tests to determine if what you see in the data (e.g., differences in the means) are reliable (replicable), and if they are likely caused by the independent variables, and not due to random effects
  - e.g., t-test to compare two means
  - e.g., ANOVA (Analysis of Variance) to compare several means
  - e.g., test “significance level” of a correlation between two variables

# Means Not Always Perfect

**Experiment 1**

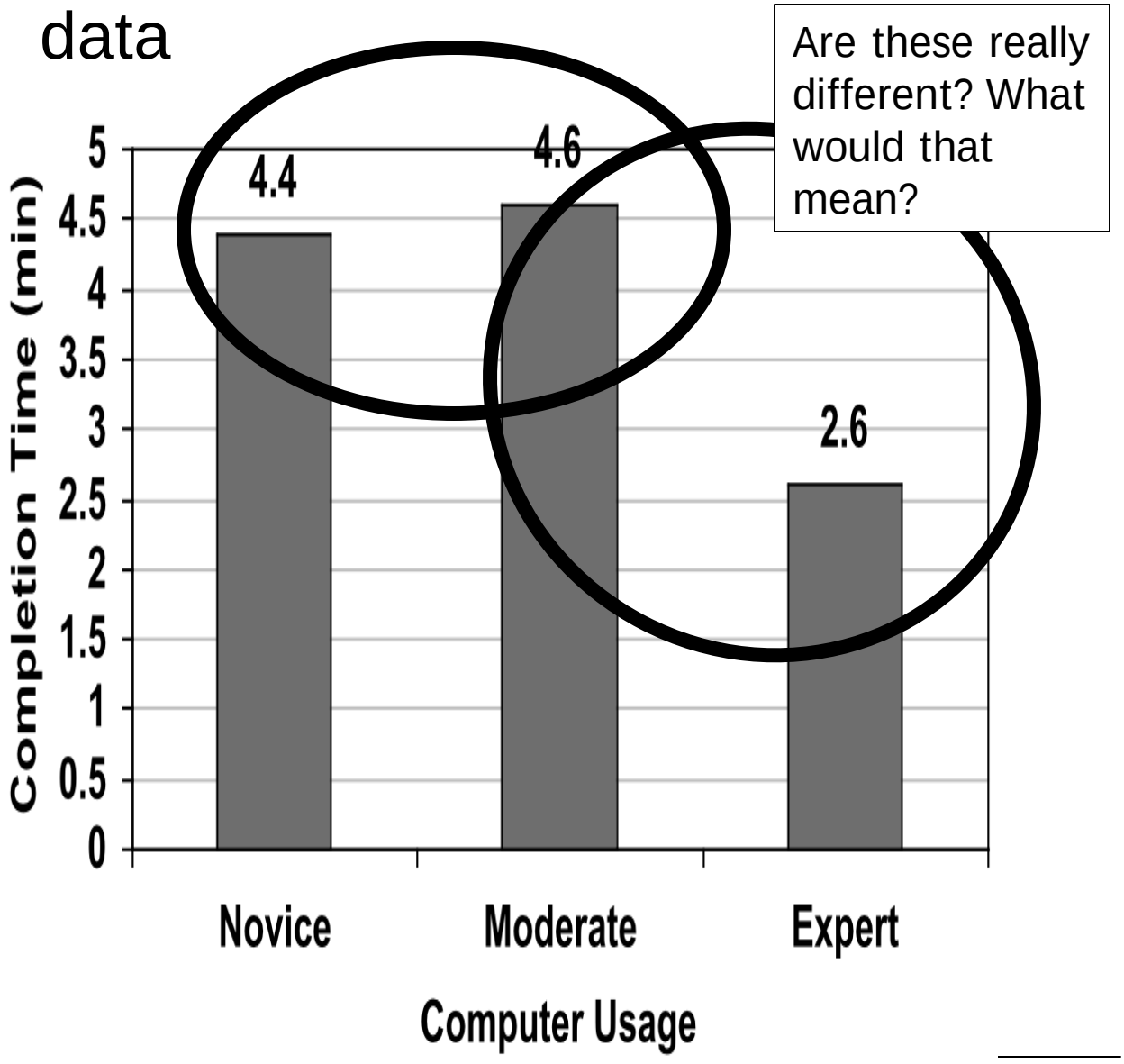
| Group 1 | Group 2 |
|---|---|
| Mean: 7 | Mean: 10 |
| 1,10,10 | 3,6,21 |

**Experiment 2**

| Group 1 | Group 2 |
|---|---|
| Mean: 7 | Mean: 10 |
| 6,7,8 | 8,11,11 |

# Inferential Stats and the Data

- Ask diagnostic questions about the data



# Hypothesis Testing

- Recall: We set up a “null hypothesis”
  - e.g., there should be no difference between the completion times of the three groups
  - Or, $H_0$: $Time_{Novice} = Time_{Moderate} = Time_{Expert}$

- Our real hypothesis was, say, that experts should perform more quickly than novices

# Hypothesis Testing

- "Significance level" (p):
  - The probability that your null hypothesis was wrong, simply by chance
  - Can also think of this as the probability that your "real" hypothesis (not the null), is wrong
  - The cutoff or threshold level of p ("alpha" level) is often set at 0.05, or 5% of the time you'll get the result you saw, just by chance
  - e.g. If your statistical t-test (testing the difference between two means) returns a t-value of t=4.5, and a p-value of p=.01, the difference between the means is statistically significant

# Errors

- Errors in analysis do occur
- Main Types:
  - Type I/False positive - You conclude there is a difference, when in fact there isn't
  - Type II/False negative - You conclude there is no different when there is
  - Dreaded Type III

# Drawing Conclusions

- Make your conclusions based on the descriptive stats, but back them up with inferential stats
  - e.g., “The expert group performed faster than the novice group $t(1,34) = 4.6$, $p > .01$.”
- Translate the stats into words that regular people can understand
  - e.g., “Thus, those who have computer experience will be able to perform better, right from the beginning…”

# Feeding Back Into Design

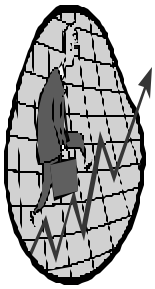
- Your study, was designed to yield information you can use to redesign your interface
- What were the conclusions you reached?
- How can you improve on the design?
- What are quantitative benefits of the redesign?
  - e.g., 2 minutes saved per transaction, which means 24% increase in production, or $45,000,000 per year in increased profit
- What are qualitative, less tangible benefit(s)?
  - e.g., workers will be less bored, less tired, and therefore more interested --> better cust. service

# Usability Specifications

"Is it good enough…

…to stop working on it?

…to get paid?"



- Quantitative usability goals, used a guide for knowing when interface is "good enough"
- Should be established as early as possible
  - Generally a large part of the Requirements Specifications at the center of a design contract
  - Evaluation is often used to demonstrate the design meets certain requirements (and so the designer/developer should get paid)
  - Often driven by competition's usability, features, or performance

# Measurement Process

- “If you can’t measure it, you can’t manage it”

- Need to keep gathering data on each iterative evaluation and refinement
- Compare benchmark task performance to specified levels
- Know when to get it out the door!

# What is Included?

- Common usability attributes that are often captured in usability specs:
  - Initial performance
  - Long-term performance
  - Learnability
  - Retainability
  - Advanced feature usage
  - First impression
  - Long-term user satisfaction

Quantitative

# Assessment Technique

## How will you judge whether your design meets the criteria?

| Usability attribute | Measure instrum. | Value to be meas. | Current level | Worst perf. level | Planned target level | Best poss level |
|---|---|---|---|---|---|---|
| **Initial perf** | Benchmk task | Length of time to successfully add appointment on the first trial | 15 secs (manual) | 30 secs | 20 secs | 10 secs |
| **First impression** | Quest | -2..2 | ?? | 0 | 0.75 | 1.5 |

# Fields

- Measuring Instrument
  - Questionnaires, Benchmark tasks
- Value to be measured
  - Time to complete task
  - Number of percentage of errors
  - Percent of task completed in given time
  - Ratio of successes to failures
  - Number of commands used
  - Frequency of help usage
- Target level
  - Often established by comparison with competing system or non-computer based task

# Summary

- Usability specs can be useful in tracking the effectiveness of redesign efforts
- They are often part of a contract
- Designers can set their own usability specs, even if the project does not specify them in advance
- Know when it is good enough, and be confident to move on to the next project