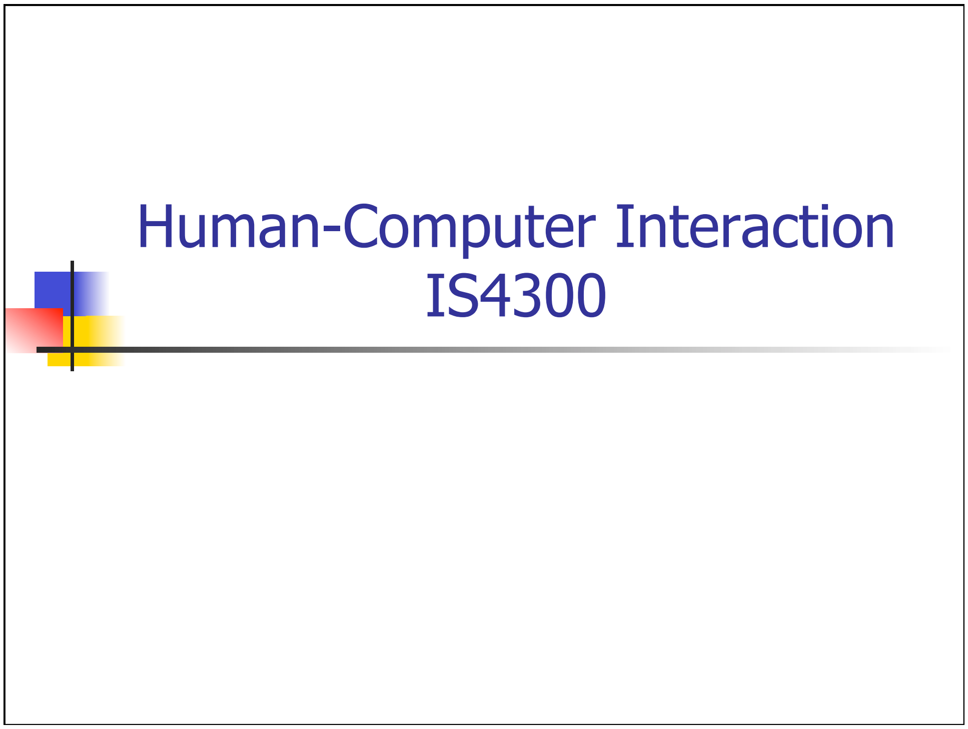
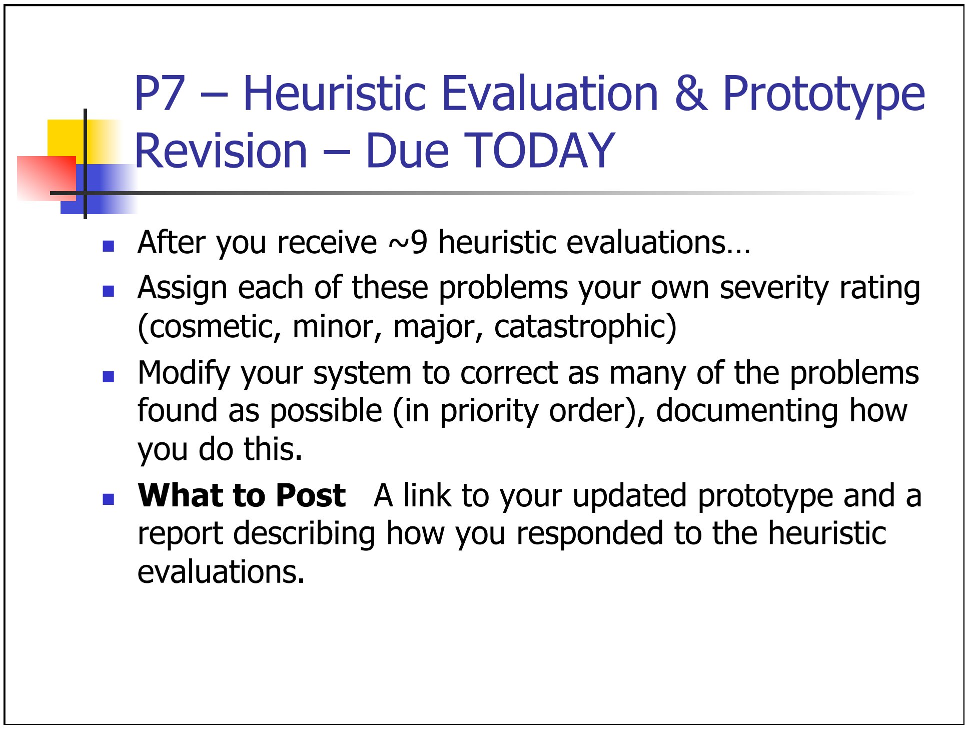


Human-Computer Interaction
IS4300



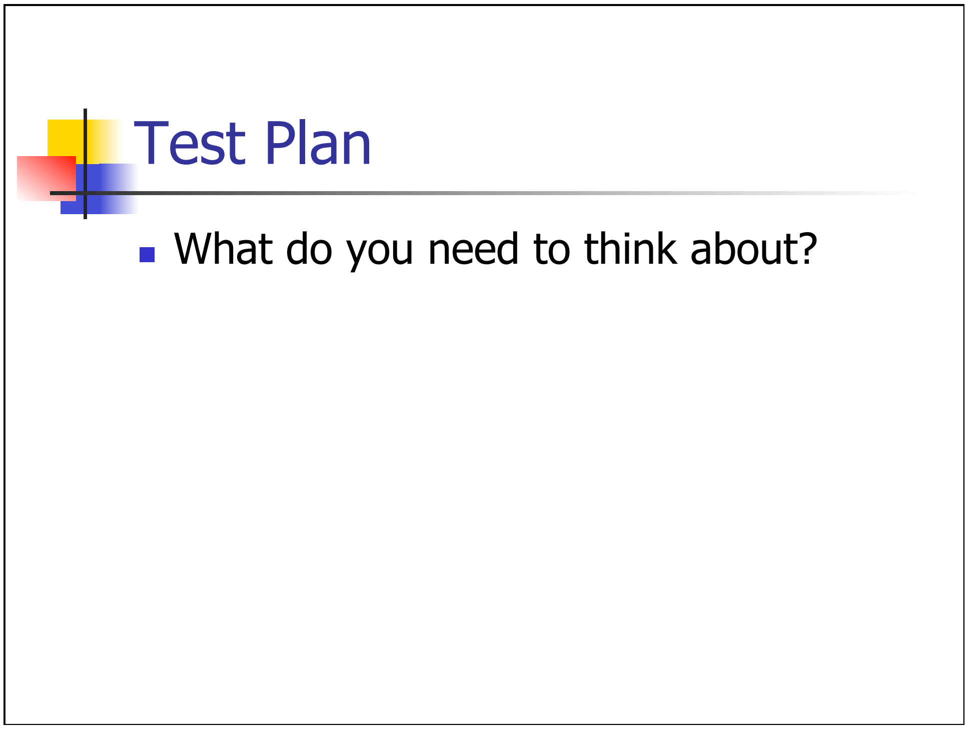
P7 – Heuristic Evaluation & Prototype Revision – Due TODAY

- After you receive ~9 heuristic evaluations…
- Assign each of these problems your own severity rating (cosmetic, minor, major, catastrophic)
- Modify your system to correct as many of the problems found as possible (in priority order), documenting how you do this.
- **What to Post** A link to your updated prototype and a report describing how you responded to the heuristic evaluations.

P8 – Finish Project & Do User Testing – Due 12/7

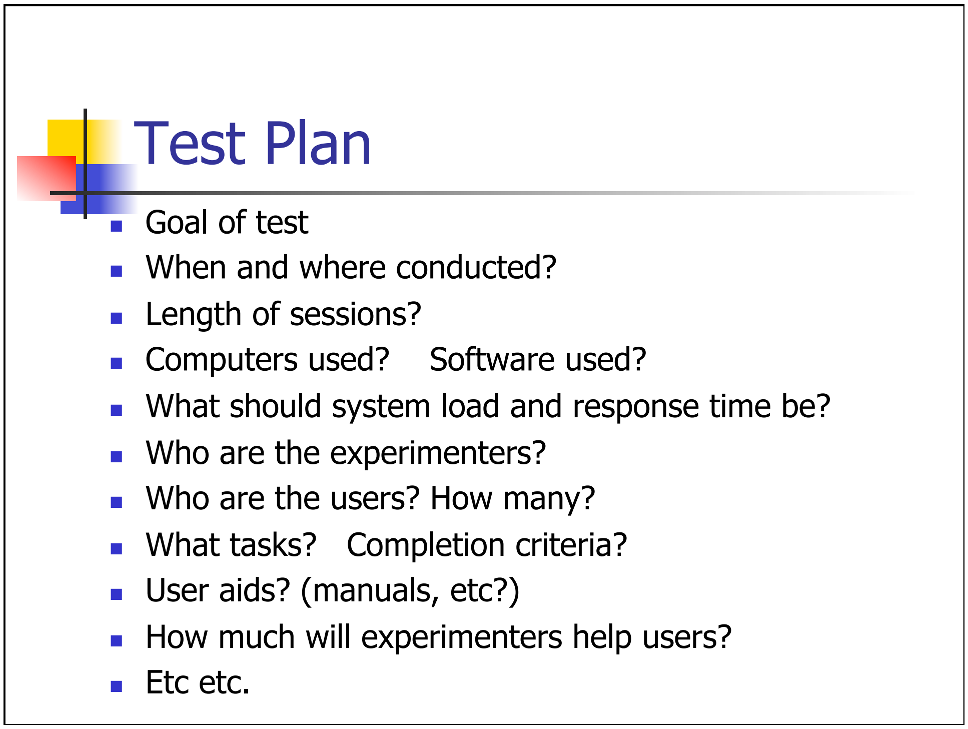
- Complete enough of your implementation to support user testing
  - Should be fully functional unless you have a compelling rationale
- Complete user testing
  - Exactly as you did in Paper Prototyping, but with your software prototype
  - 3+ users (not classmates), 3+ tasks
  - Briefing
  - Can demo system on additional task first
- Redesign
  - Sort severity problems by severity
  - Address as many as possible
- Document everything
- Post
  - Final software prototype
  - Report

Review:
Conducting Usability Studies



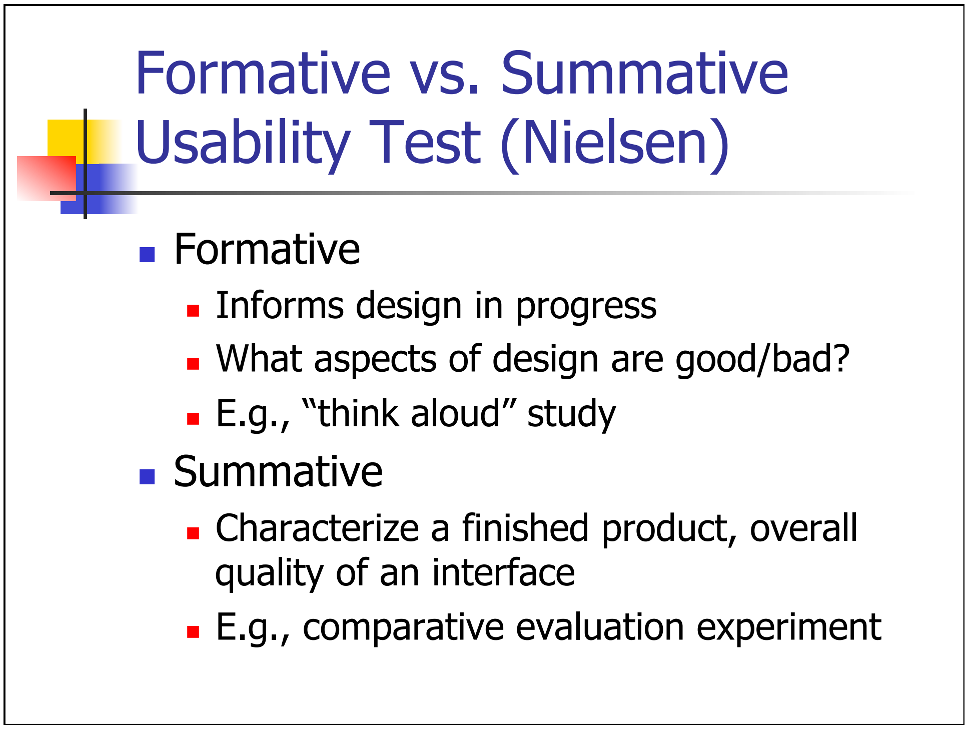
Test Plan

- What do you need to think about?



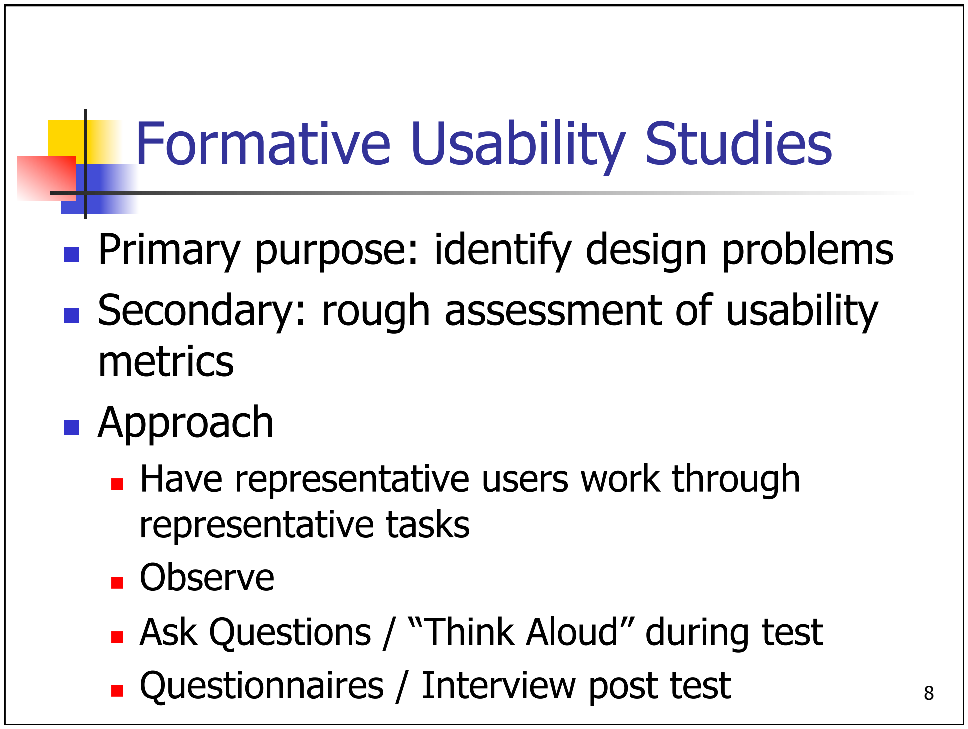
Test Plan

- Goal of test
- When and where conducted?
- Length of sessions?
- Computers used?   Software used?
- What should system load and response time be?
- Who are the experimenters?
- Who are the users? How many?
- What tasks?   Completion criteria?
- User aids? (manuals, etc?)
- How much will experimenters help users?
- Etc etc.



Formative vs. Summative Usability Test (Nielsen)

- Formative
  - Informs design in progress
  - What aspects of design are good/bad?
  - E.g., “think aloud” study
- Summative
  - Characterize a finished product, overall quality of an interface
  - E.g., comparative evaluation experiment



Formative Usability Studies

- Primary purpose: identify design problems
- Secondary: rough assessment of usability metrics
- Approach
  - Have representative users work through representative tasks
  - Observe
  - Ask Questions / “Think Aloud” during test
  - Questionnaires / Interview post test

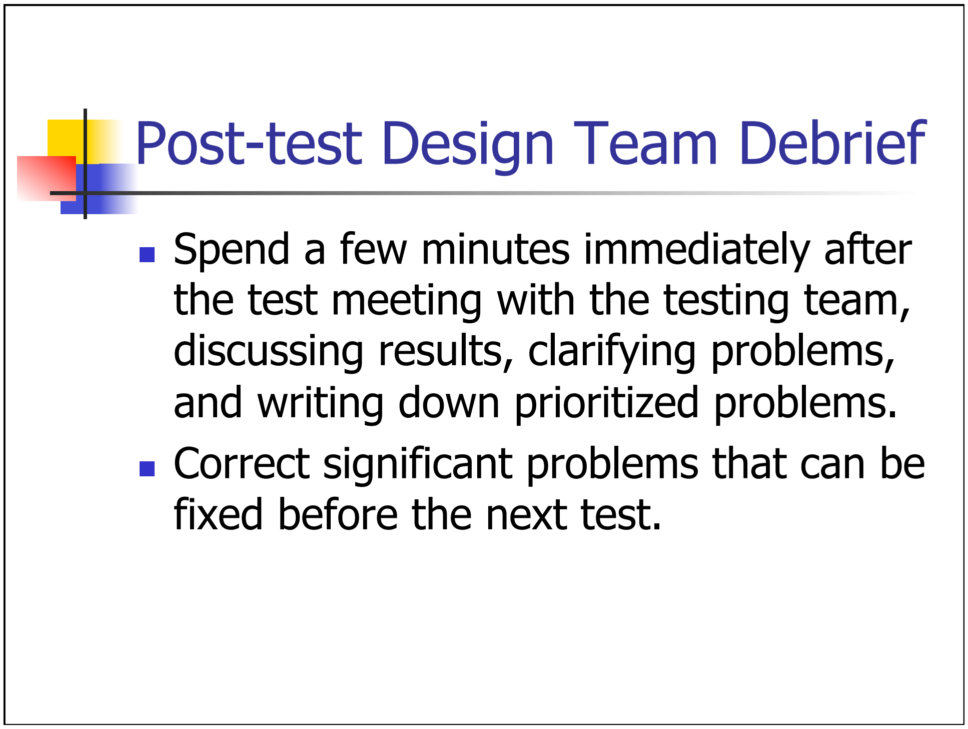
8

Facilitator – during test

- Encourage questions but don’t answer them
- Use user’s vocabulary
- Use open-ended questions
  - “What will that do?”
  - “What are you trying to do right now?”
  - “What are you thinking?”
  - “Tell me more about that.”
- Watch for “hmm”, “ah”, “oh”, “oops”, furrowed brow, etc. - ask what’s going on.
- Make changes during test or between tests if necessary
- Take a break if something goes wrong

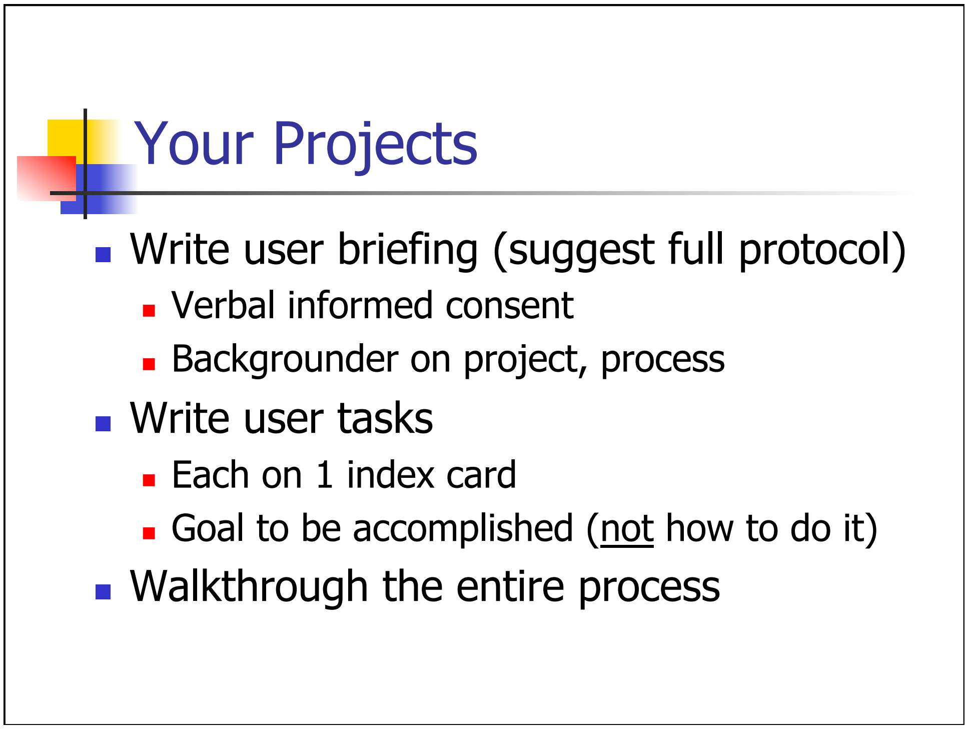
Additional questions: Think-Aloud and Offering Help

- Using Cognitive Walkthrough Questions
  - “Is there anything there that tells you what to do next?”
  - “Is there a choice on the screen that lines up with what you want to do? If so, which one?”
  - “Now that you’ve tried it, has it done what you wanted it to do?”



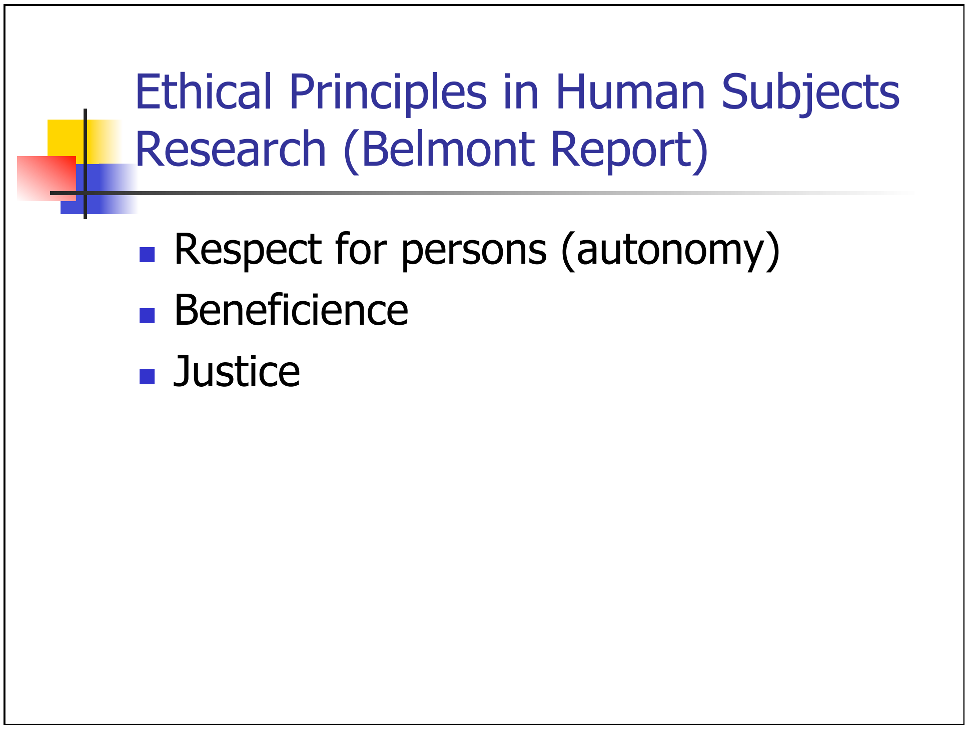
Post-test Design Team Debrief

- Spend a few minutes immediately after the test meeting with the testing team, discussing results, clarifying problems, and writing down prioritized problems.
- Correct significant problems that can be fixed before the next test.



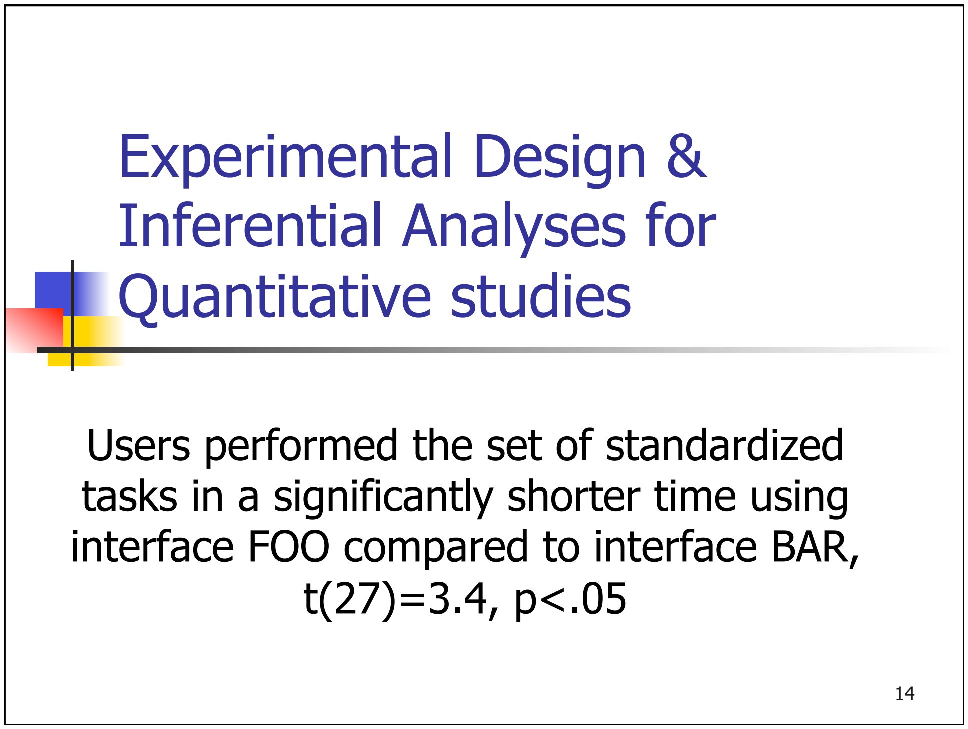
Your Projects

- Write user briefing (suggest full protocol)
  - Verbal informed consent
  - Backgrounder on project, process
- Write user tasks
  - Each on 1 index card
  - Goal to be accomplished (not how to do it)
- Walkthrough the entire process



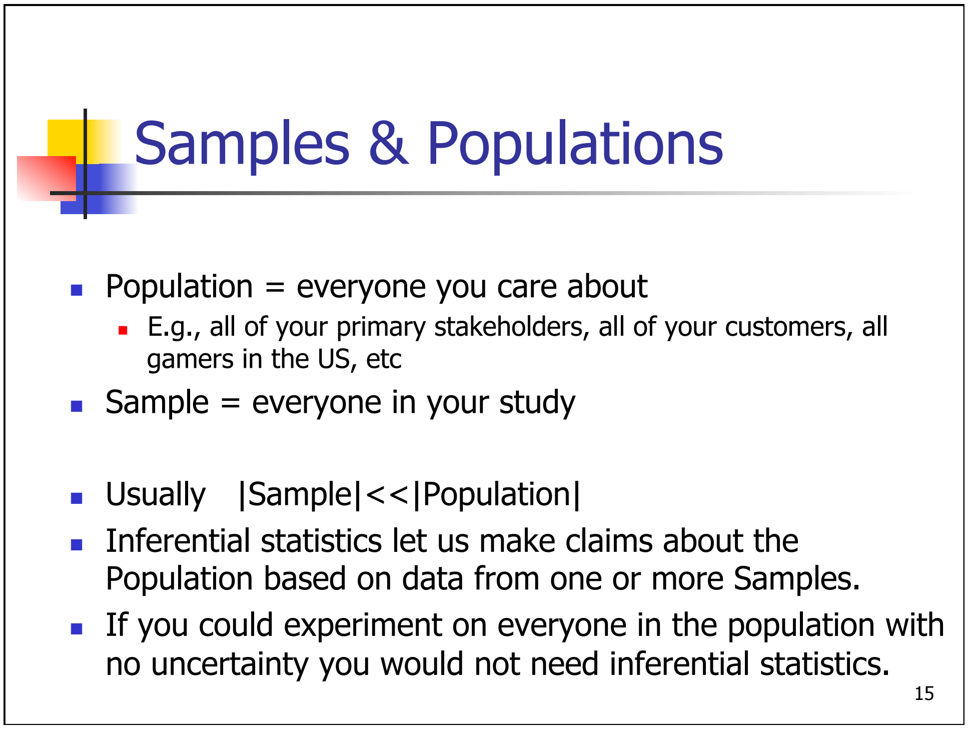
Ethical Principles in Human Subjects Research (Belmont Report)

- Respect for persons (autonomy)
- Beneficience
- Justice



Experimental Design & Inferential Analyses for Quantitative studies

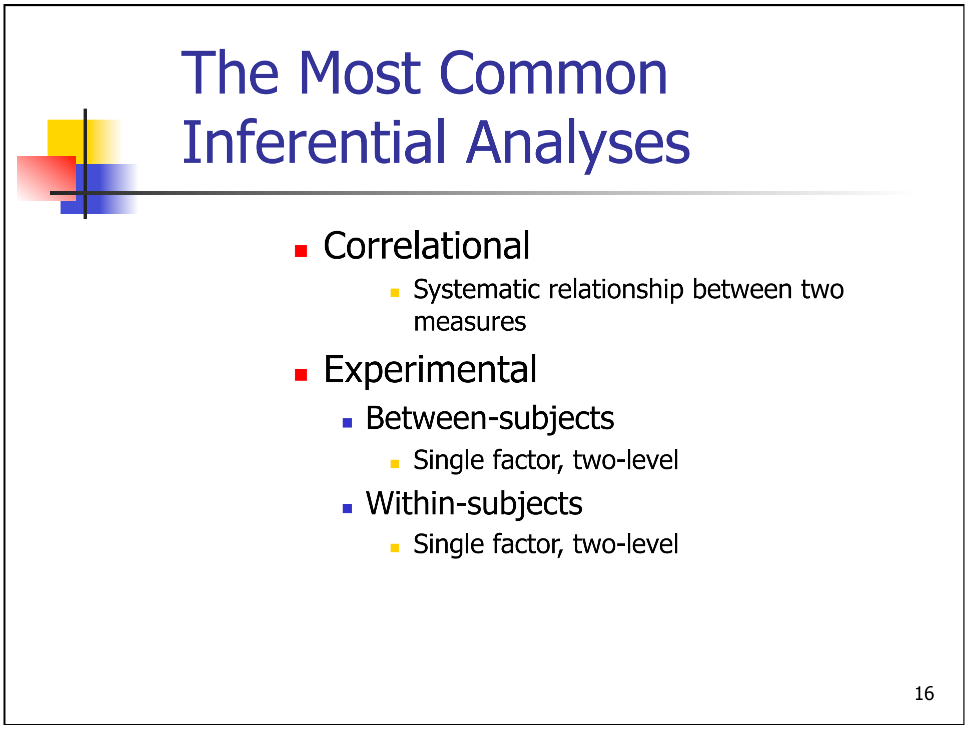
Users performed the set of standardized tasks in a significantly shorter time using interface FOO compared to interface BAR, $t(27)=3.4$, $p<.05$



Samples & Populations

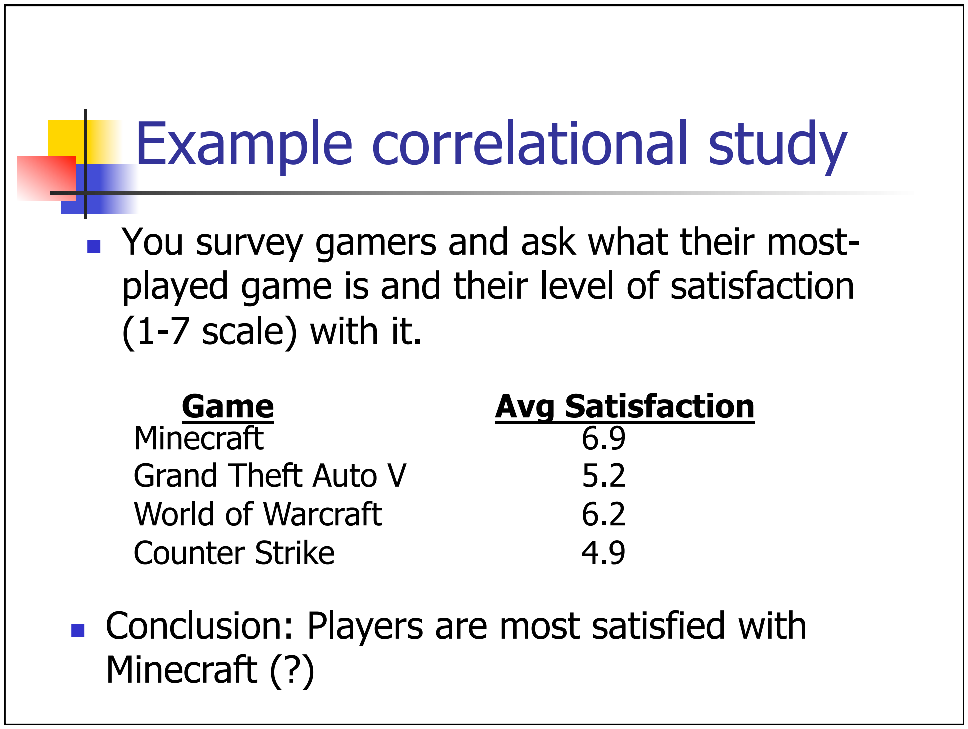
- Population = everyone you care about
  - E.g., all of your primary stakeholders, all of your customers, all gamers in the US, etc
- Sample = everyone in your study

- Usually |Sample|<<|Population|
- Inferential statistics let us make claims about the Population based on data from one or more Samples.
- If you could experiment on everyone in the population with no uncertainty you would not need inferential statistics.



The Most Common Inferential Analyses

- Correlational
  - Systematic relationship between two measures
- Experimental
  - Between-subjects
    - Single factor, two-level
  - Within-subjects
    - Single factor, two-level

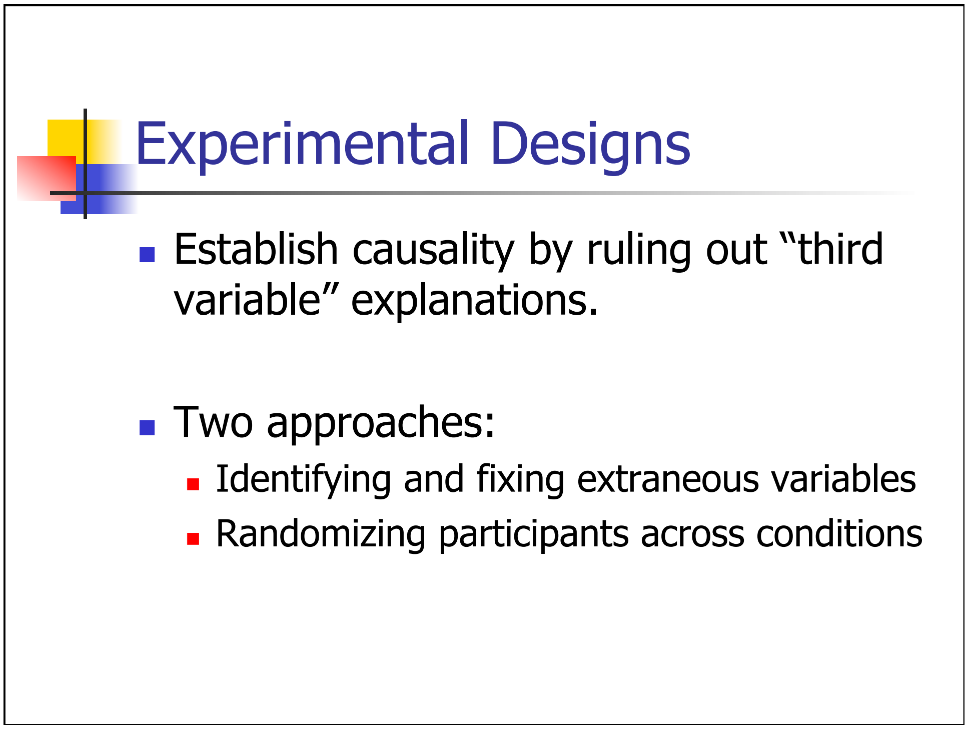


Example correlational study

- You survey gamers and ask what their most-played game is and their level of satisfaction (1-7 scale) with it.

Game	**Avg Satisfaction**
Minecraft	6.9
Grand Theft Auto V	5.2
World of Warcraft	6.2
Counter Strike	4.9

- Conclusion: Players are most satisfied with Minecraft (?)



Experimental Designs

- Establish causality by ruling out "third variable" explanations.
- Two approaches:
  - Identifying and fixing extraneous variables
  - Randomizing participants across conditions

Typical case

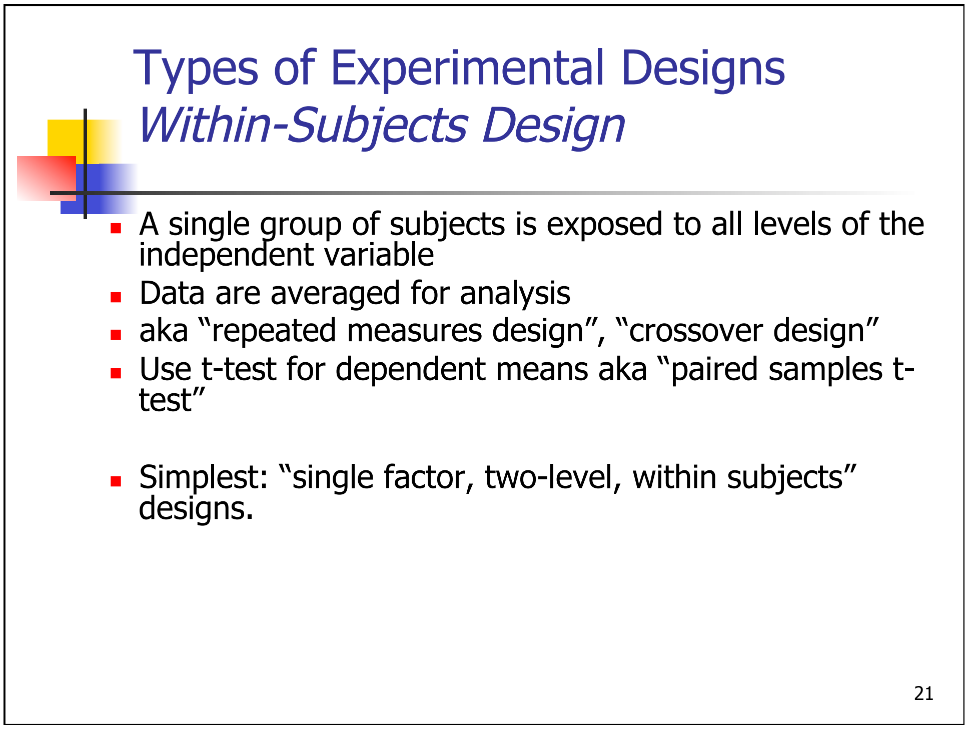
- You are trying to demonstrate there is a difference between two designs/ interfaces/systems based on a usability metrics
  - E.g., performance with interface FOO vs. performance with interface BAR
  - E.g., Satisfaction with Minecraft vs. World of Warcraft

19

Types of Experimental Designs *Between-Subjects Design*

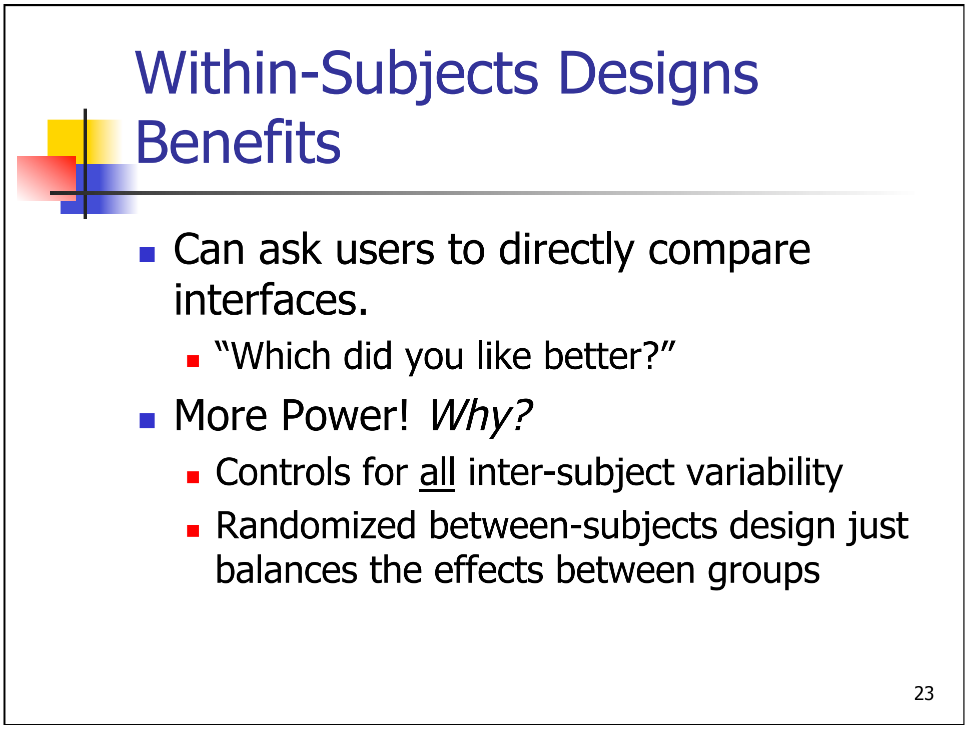
-
  - Different groups of subjects are randomly assigned to the levels of your independent variable
  - Data are averaged for analysis
  - If interval or ratio measures and approximately normal, use t-test for independent means
  - Simplest: "single factor, two-level, between subjects" designs.

20



Types of Experimental Designs *Within-Subjects Design*

- A single group of subjects is exposed to all levels of the independent variable
- Data are averaged for analysis
- aka “repeated measures design”, “crossover design”
- Use t-test for dependent means aka “paired samples t-test”
- Simplest: “single factor, two-level, within subjects” designs.



Within-Subjects Designs Benefits

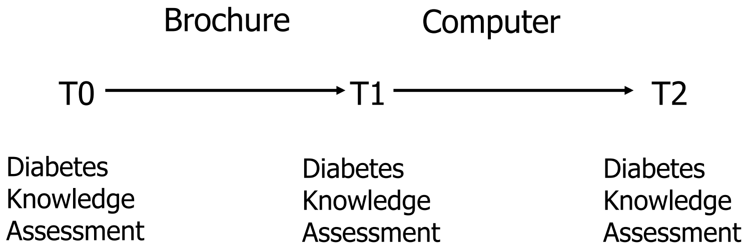
- Can ask users to directly compare interfaces.
  - “Which did you like better?”
- More Power! *Why?*
  - Controls for all inter-subject variability
  - Randomized between-subjects design just balances the effects between groups

Within-Subjects Designs Disadvantages

- More demanding on subjects, especially in complex designs
- Subject attrition is a problem
- *Carryover effects:* Exposure to a previous treatment affects performance in a subsequent treatment

Carryover Example

- Embodied Conversational Agents to Promote Health Literacy for Older Adults



Brochure

Computer

T0 → T1 → T2

Diabetes Knowledge Assessment

Diabetes Knowledge Assessment

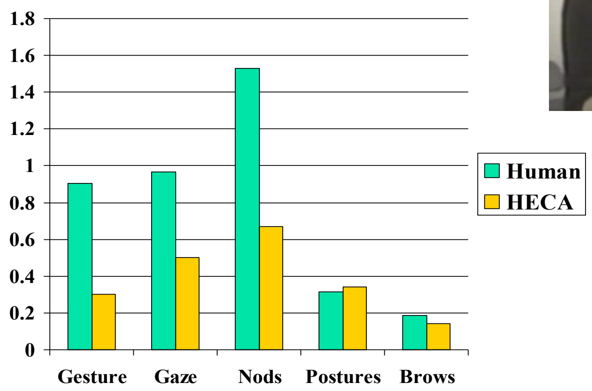
Diabetes Knowledge Assessment

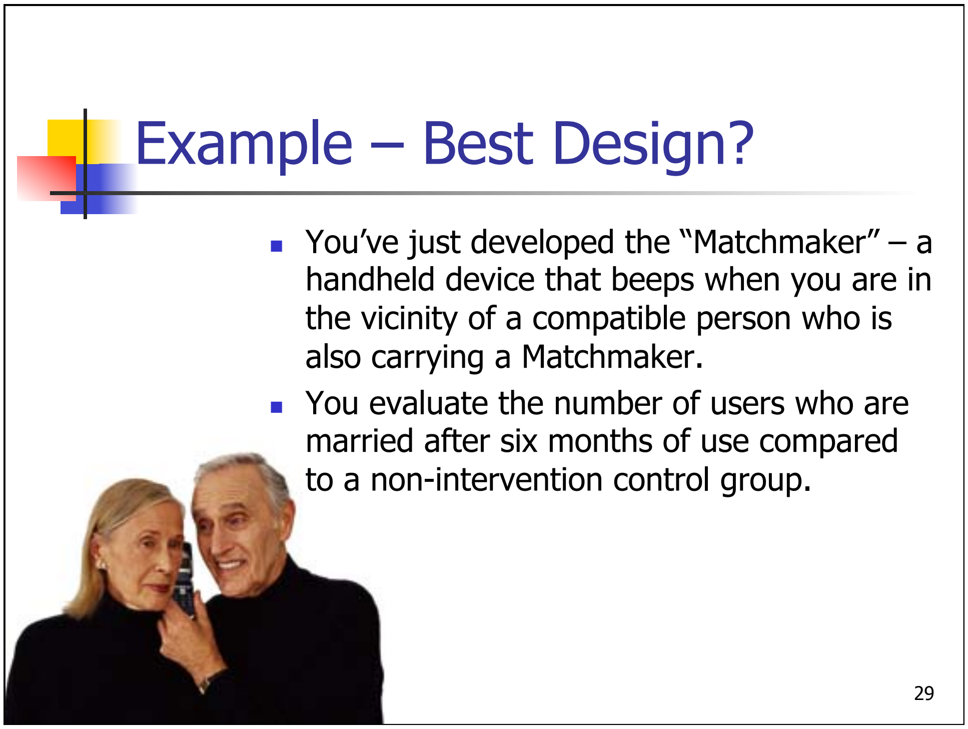
Some Sources of Carryover

- *Learning*
  - Learning a task in the first treatment may affect performance in the second
- *Fatigue*
  - Fatigue from earlier treatments may affect performance in later treatments
- *Habituation*
  - Repeated exposure to a stimulus may lead to unresponsiveness to that stimulus
- *Sensitization*
  - Exposure to a stimulus may make a subject respond more strongly to another
- *Contrast*
  - Subjects may compare treatments, which may affect behavior

27

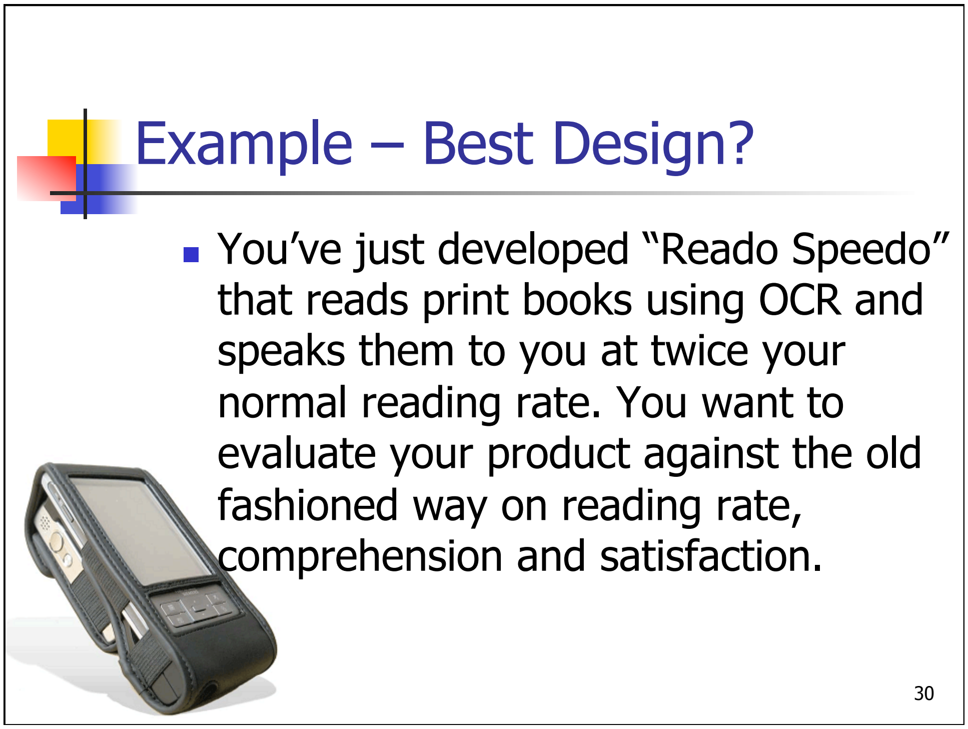
Example Study – Best design? Handheld ECAs





Example – Best Design?

- You’ve just developed the “Matchmaker” – a handheld device that beeps when you are in the vicinity of a compatible person who is also carrying a Matchmaker.
- You evaluate the number of users who are married after six months of use compared to a non-intervention control group.



Example – Best Design?

- You’ve just developed “Reado Speedo” that reads print books using OCR and speaks them to you at twice your normal reading rate. You want to evaluate your product against the old fashioned way on reading rate, comprehension and satisfaction.

Example – Best Design?

- You've developed a new web-based help system for your email client. You want to compare your system to the old printed manual.

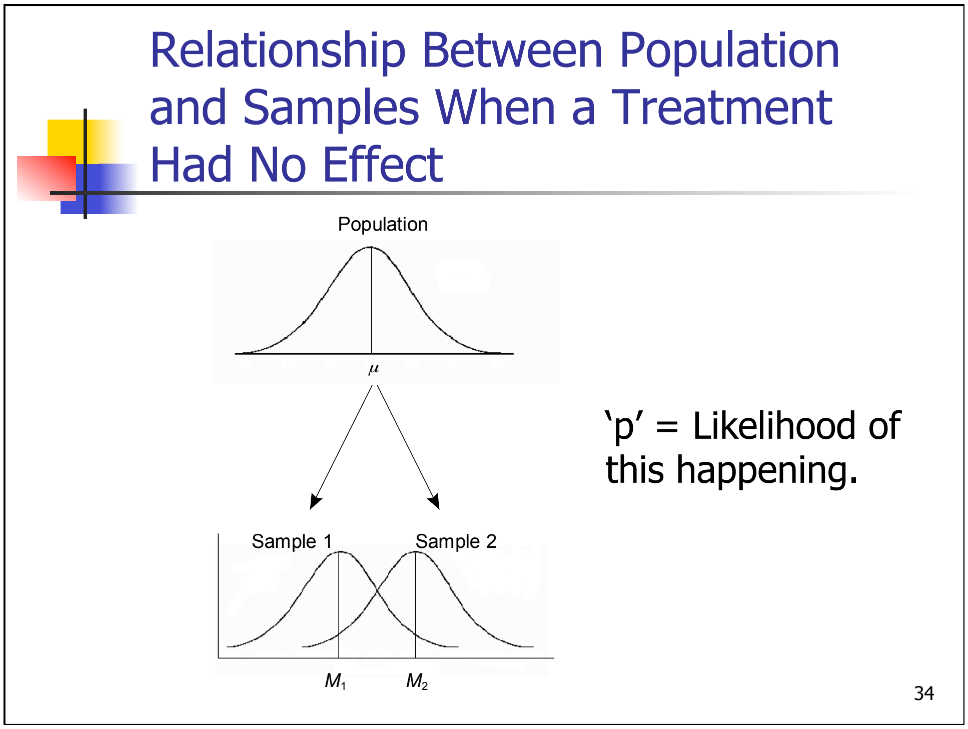
31

Type of Errors in Inferential Statistics

Research Hypothesis: There is a difference
(e.g., FOO better than BAR)

	"The Truth"	
Your study…	**No diff**	**Diff**
Conclude diff	**Type I Error**	**Correct Decision**
Conclude no diff	**Correct Decision**	**Type II Error**

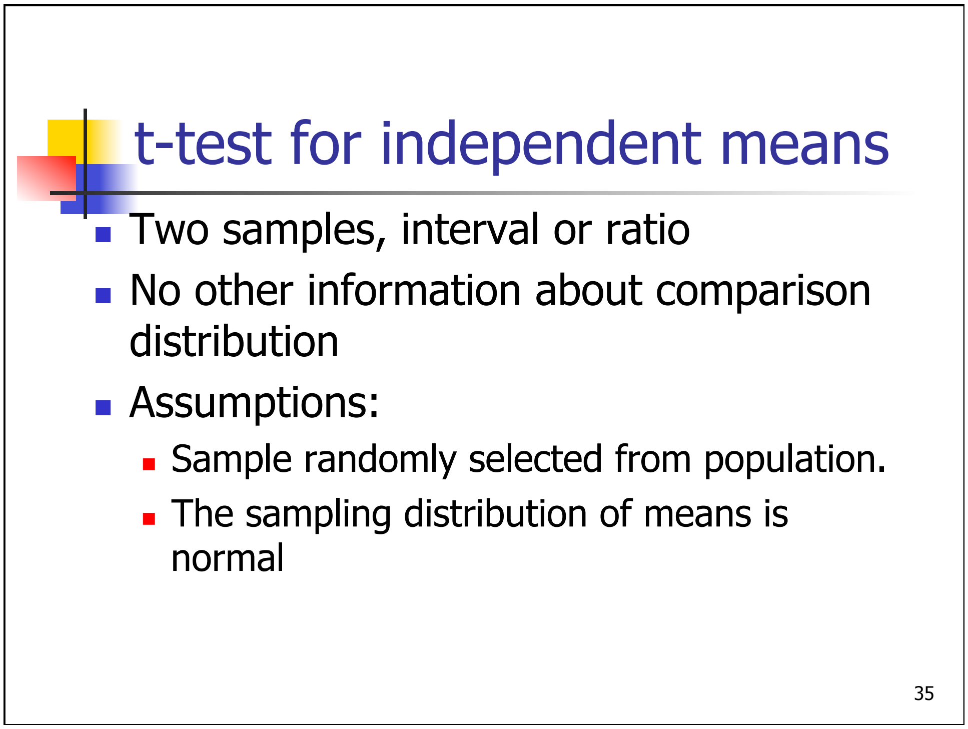
'p' = Probability of Type I Error
The likelihood the difference observed is not real – its only due to noise / random error.



Relationship Between Population and Samples When a Treatment Had No Effect

'p' = Likelihood of this happening.

34

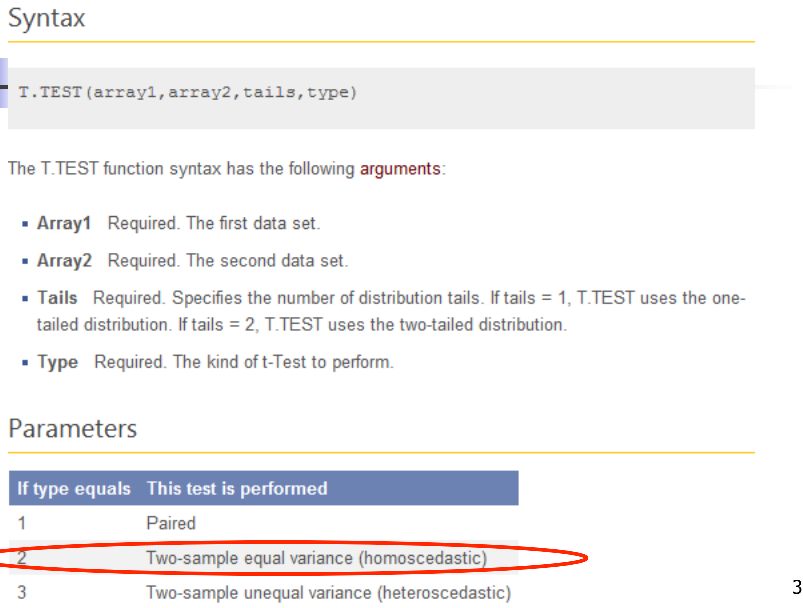


t-test for independent means

- Two samples, interval or ratio
- No other information about comparison distribution
- Assumptions:
  - Sample randomly selected from population.
  - The sampling distribution of means is normal

35

Excel T.TEST, returns ‘p’



Syntax

```
T.TEST(array1,array2,tails,type)
```

The T.TEST function syntax has the following arguments:

- **Array1** Required. The first data set.
- **Array2** Required. The second data set.
- **Tails** Required. Specifies the number of distribution tails. If tails = 1, T.TEST uses the one-tailed distribution. If tails = 2, T.TEST uses the two-tailed distribution.
- **Type** Required. The kind of t-Test to perform.

Parameters

If type equals	This test is performed
1	Paired
2	Two-sample equal variance (homoscedastic)
3	Two-sample unequal variance (heteroscedastic)

36

t-test

- If assumptions are followed, T.TEST returns ‘p’
  - Likelihood of differences observed being due to chance, or error
  - = Probability of Type I error
- If p<threshold (conventionally 0.05), we say there is a significant difference
- If p>=threshold, we conclude nothing (experiment was inconclusive)

37

Reporting results

- Significant results, scientific articles

  t(df)=*tscore*, p<*sig*

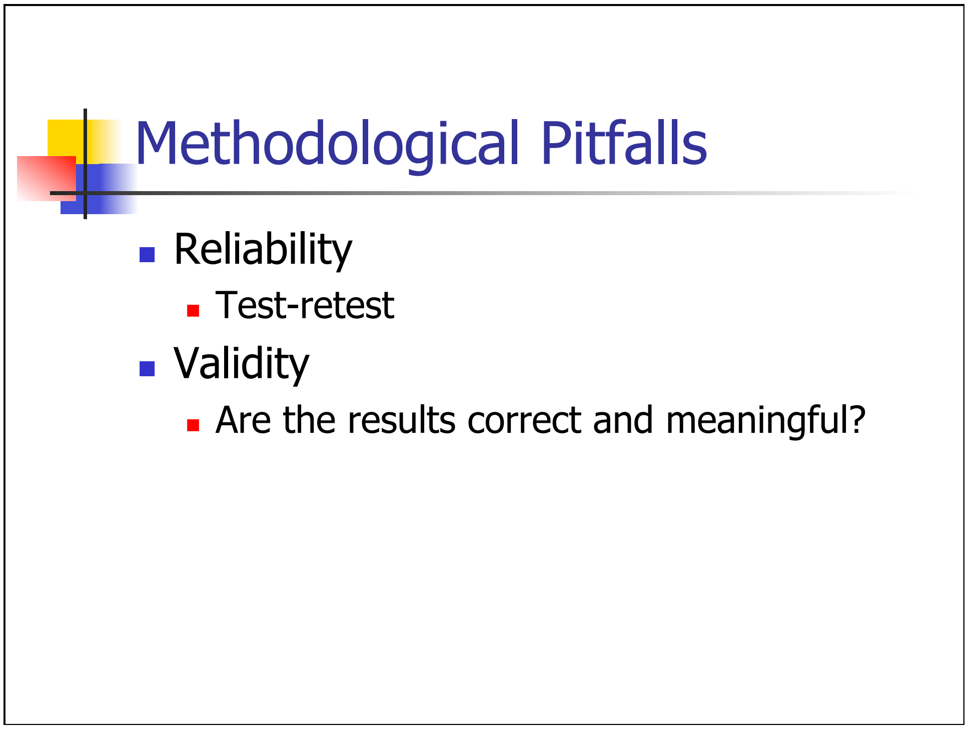
  e.g., t(38)=4.72, p<.05
- Non-significant results

  e.g., t(38)=4.72, n.s.
- df = total number of subjects - 2
- Informal usability reports:
  - t-test for independent means indicated that performance with FOO was significantly better than performance with BAR, p<.05
  - t-test for independent means for performance with FOO vs. BAR was not significant.

38

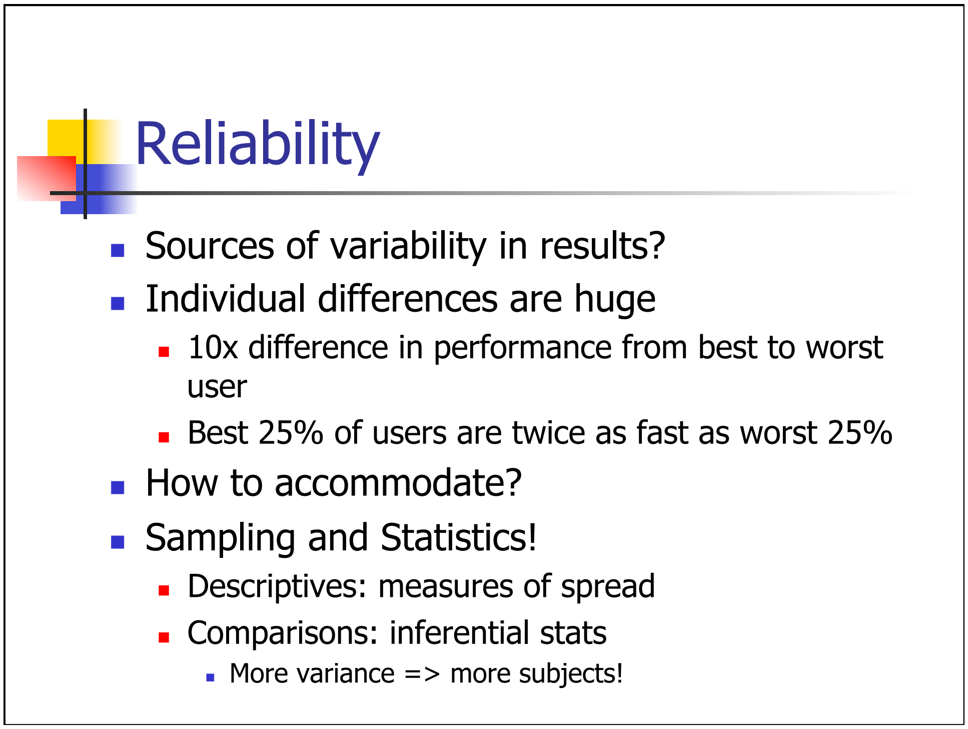
Nielsen on Usability Testing

Usability Engineering
Ch 6



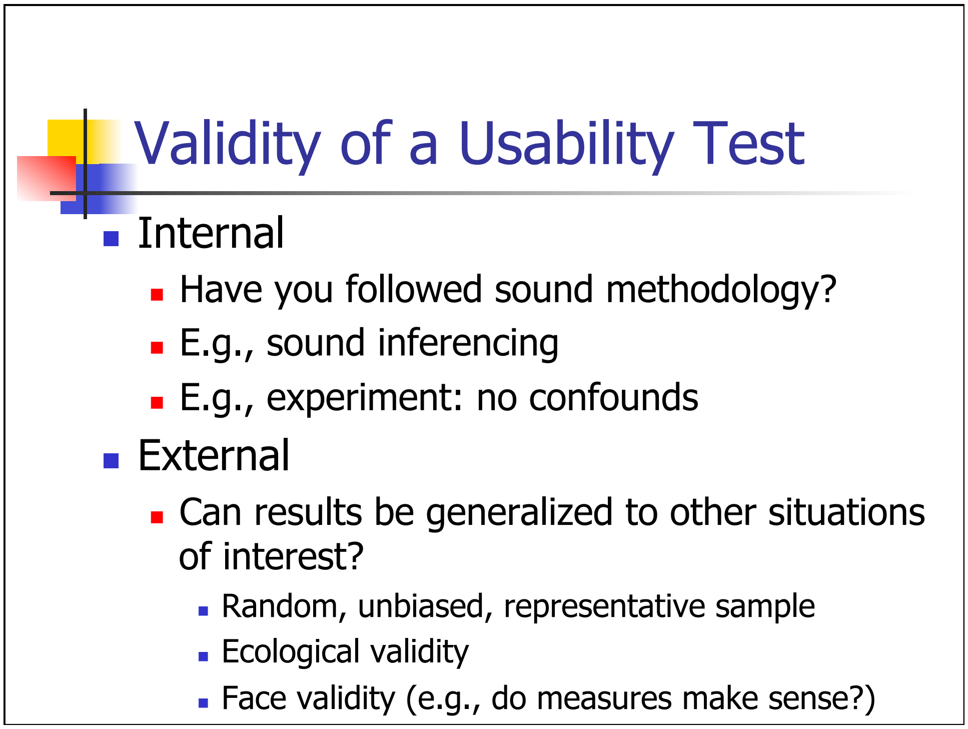
Methodological Pitfalls

- Reliability
  - Test-retest
- Validity
  - Are the results correct and meaningful?



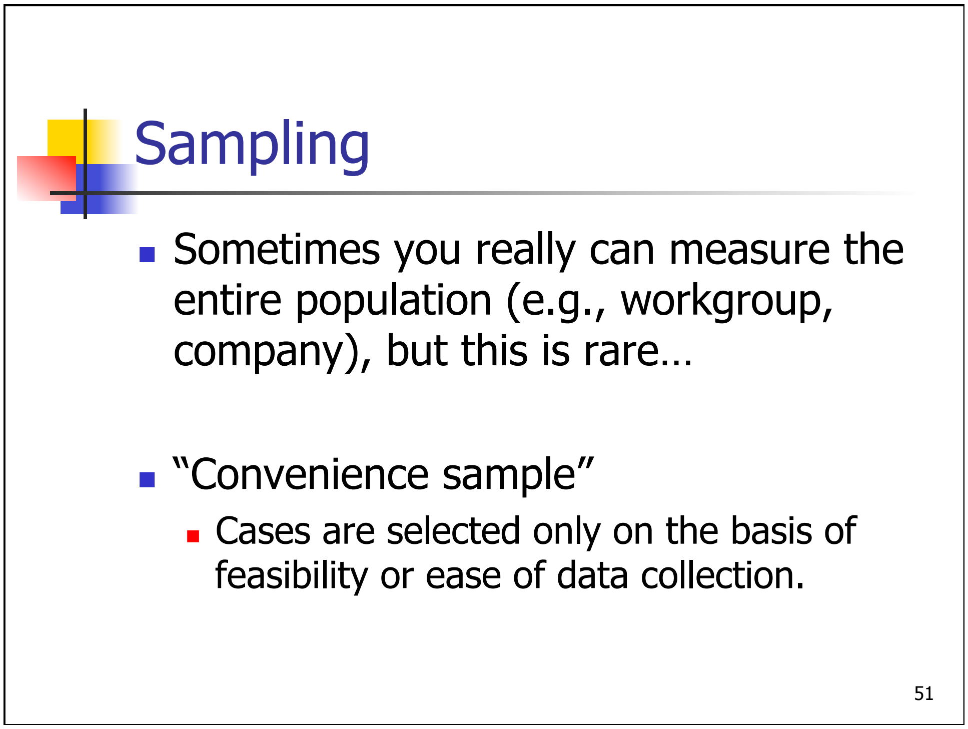
Reliability

- Sources of variability in results?
- Individual differences are huge
  - 10x difference in performance from best to worst user
  - Best 25% of users are twice as fast as worst 25%
- How to accommodate?
- Sampling and Statistics!
  - Descriptives: measures of spread
  - Comparisons: inferential stats
    - More variance => more subjects!



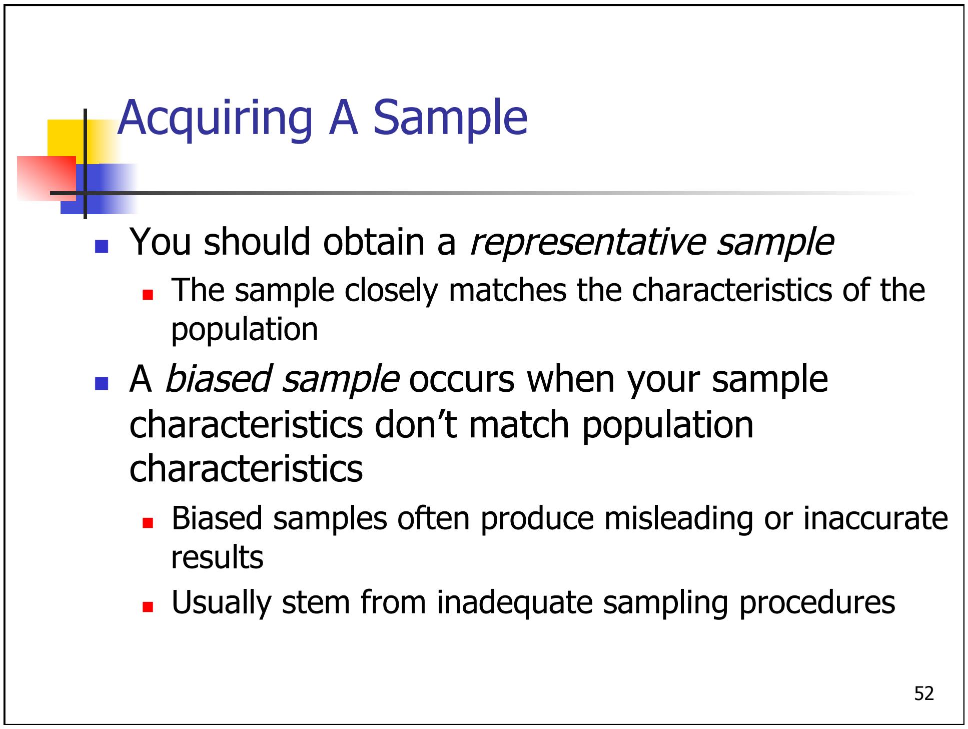
Validity of a Usability Test

- Internal
  - Have you followed sound methodology?
  - E.g., sound inferencing
  - E.g., experiment: no confounds
- External
  - Can results be generalized to other situations of interest?
    - Random, unbiased, representative sample
    - Ecological validity
    - Face validity (e.g., do measures make sense?)



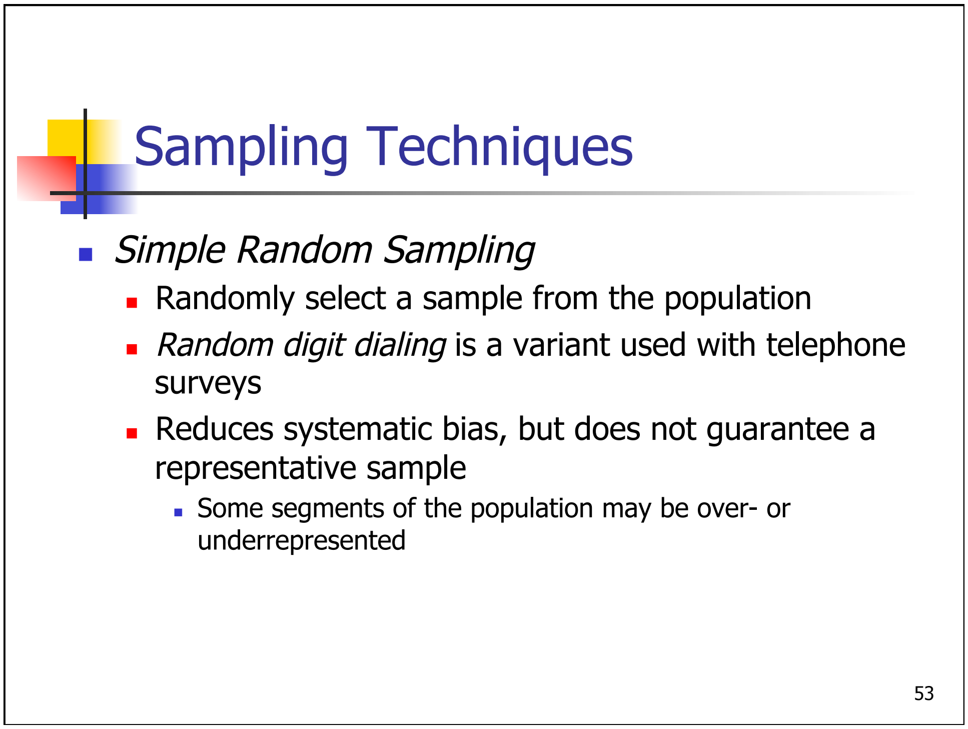
Sampling

- Sometimes you really can measure the entire population (e.g., workgroup, company), but this is rare…
- “Convenience sample”
  - Cases are selected only on the basis of feasibility or ease of data collection.



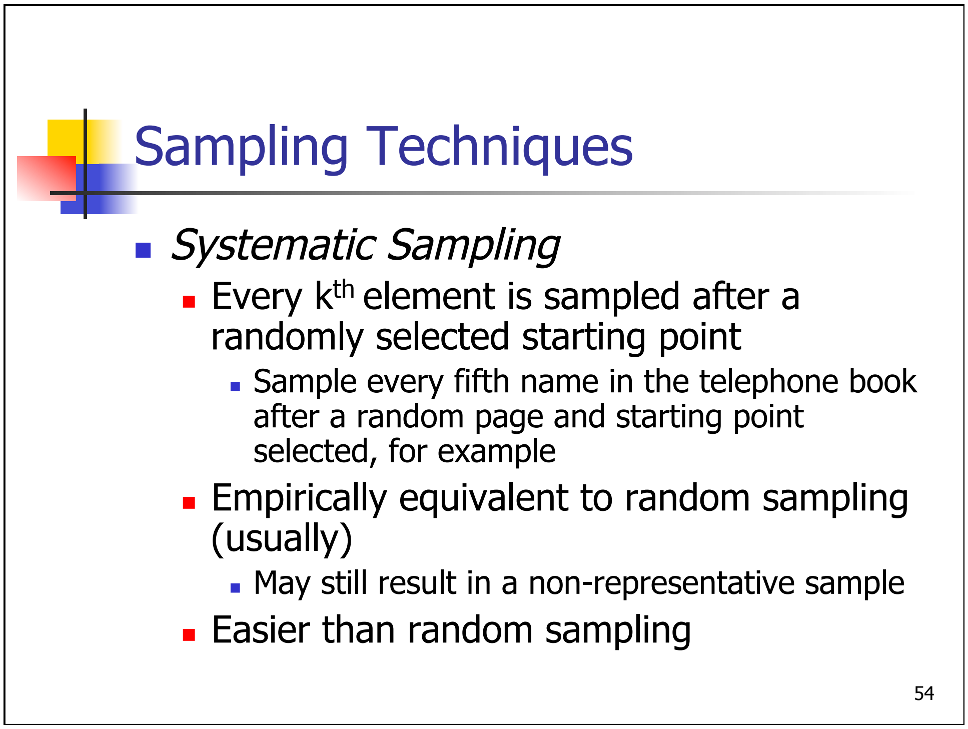
Acquiring A Sample

- You should obtain a *representative sample*
  - The sample closely matches the characteristics of the population
- A *biased sample* occurs when your sample characteristics don’t match population characteristics
  - Biased samples often produce misleading or inaccurate results
  - Usually stem from inadequate sampling procedures



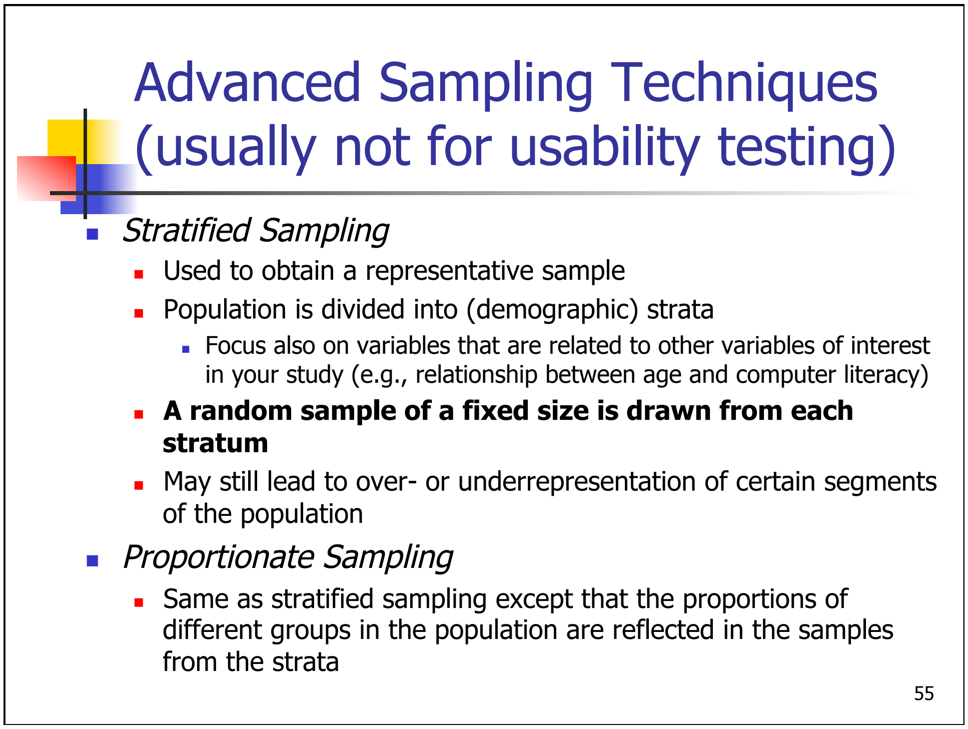
Sampling Techniques

- *Simple Random Sampling*
  - Randomly select a sample from the population
  - *Random digit dialing* is a variant used with telephone surveys
  - Reduces systematic bias, but does not guarantee a representative sample
    - Some segments of the population may be over- or underrepresented



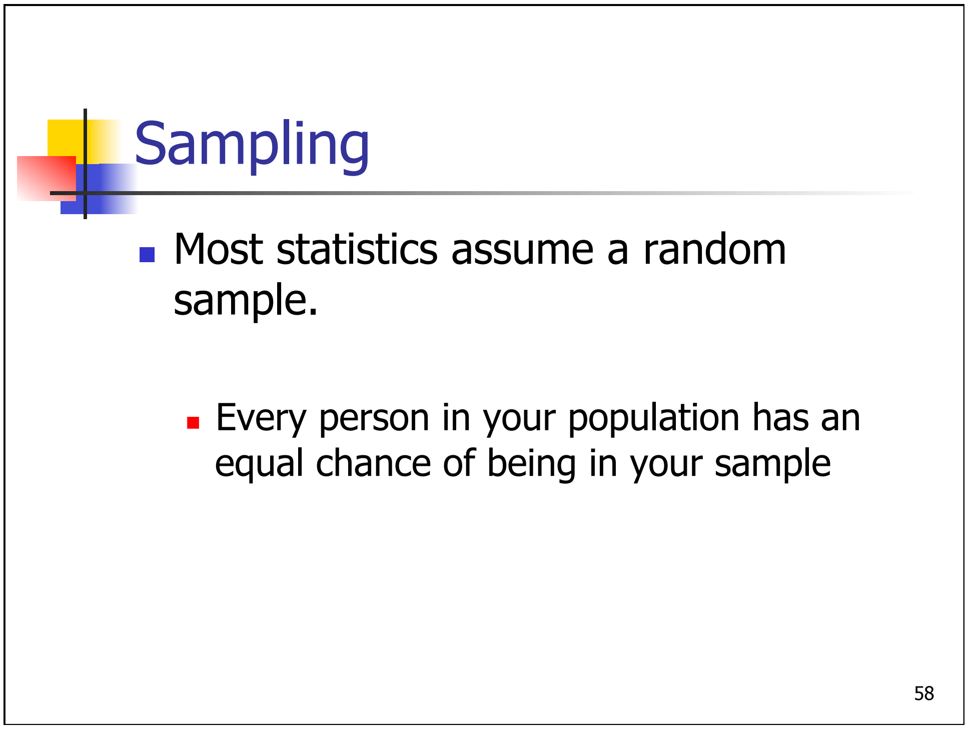
Sampling Techniques

- *Systematic Sampling*
  - Every kth element is sampled after a randomly selected starting point
    - Sample every fifth name in the telephone book after a random page and starting point selected, for example
  - Empirically equivalent to random sampling (usually)
    - May still result in a non-representative sample
  - Easier than random sampling



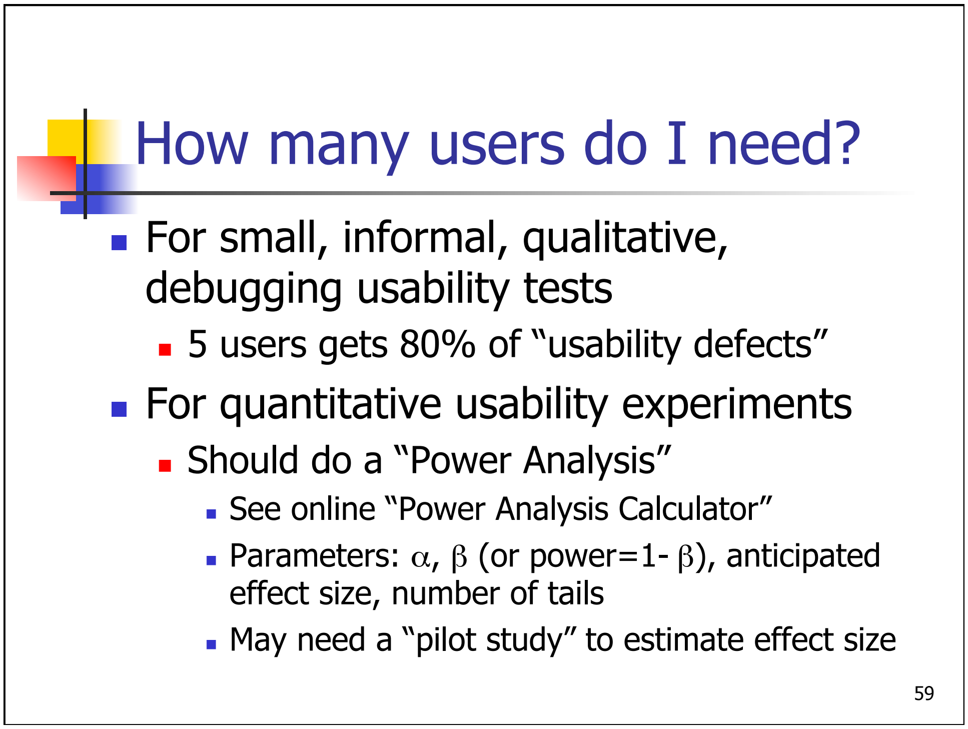
Advanced Sampling Techniques (usually not for usability testing)

- *Stratified Sampling*
  - Used to obtain a representative sample
  - Population is divided into (demographic) strata
    - Focus also on variables that are related to other variables of interest in your study (e.g., relationship between age and computer literacy)
  - **A random sample of a fixed size is drawn from each stratum**
  - May still lead to over- or underrepresentation of certain segments of the population
- *Proportionate Sampling*
  - Same as stratified sampling except that the proportions of different groups in the population are reflected in the samples from the strata



Sampling

- Most statistics assume a random sample.
  - Every person in your population has an equal chance of being in your sample



How many users do I need?

- For small, informal, qualitative, debugging usability tests
  - 5 users gets 80% of "usability defects"
- For quantitative usability experiments
  - Should do a "Power Analysis"
    - See online "Power Analysis Calculator"
    - Parameters: α, β (or power=1- β), anticipated effect size, number of tails
    - May need a "pilot study" to estimate effect size

Test Budget

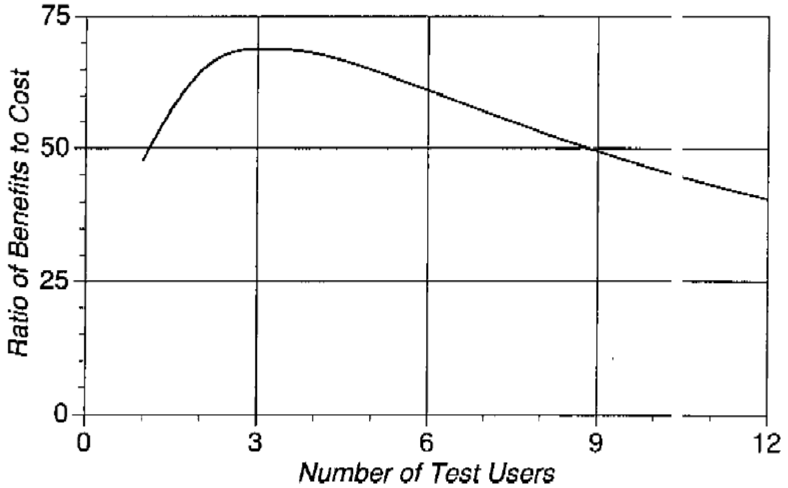
- Personnel
- Tester compensation
- Computers
- Lab
- Special equipment (e.g., gaze tracker)
- Video/audio tapes

- WAG: \$3k + \$1k/user for typical industry test
  - 1993 \$, ~+150% now)

Usability Test ROI

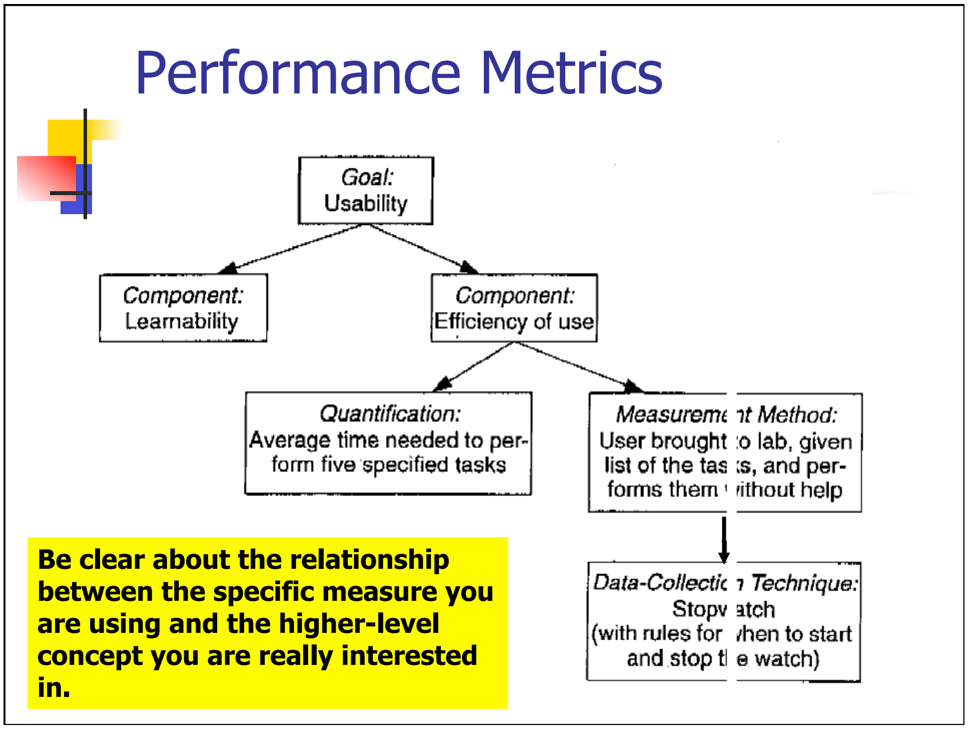
- Number of usability problems found = $N(1 - (1 - \lambda)^{i})$
  - i = number of test users
  - N = total number of usability problems
  - λ = P(finding any given problem by a given user)
- Examples
  - Value of finding a usability problem = \$15k
  - N = 41
  - $\lambda = 0.31$

Payoff ratio given these assumptions



Pilot Test

- Always run 1-2 test subjects first to debug the study protocol.
- Also used to characterize effect size to power for a larger experimental study



Performance Metrics

Goal:
Usability

Component:
Learnability

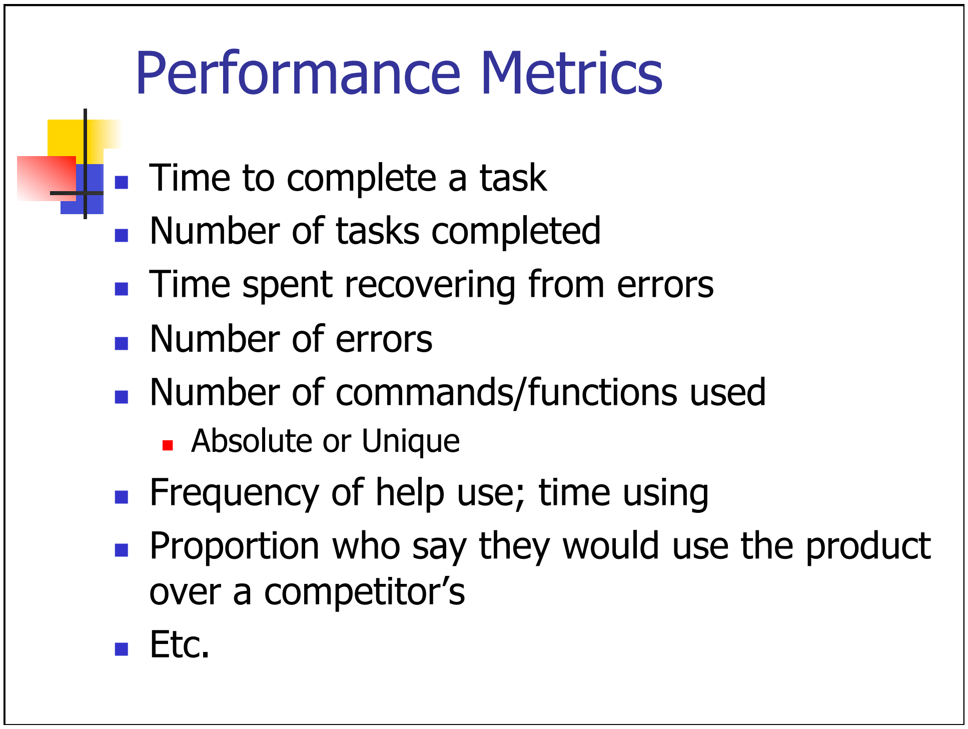
Component:
Efficiency of use

Quantification:
Average time needed to perform five specified tasks

Measurement Method:
User brought to lab, given list of the tasks, and performs them without help

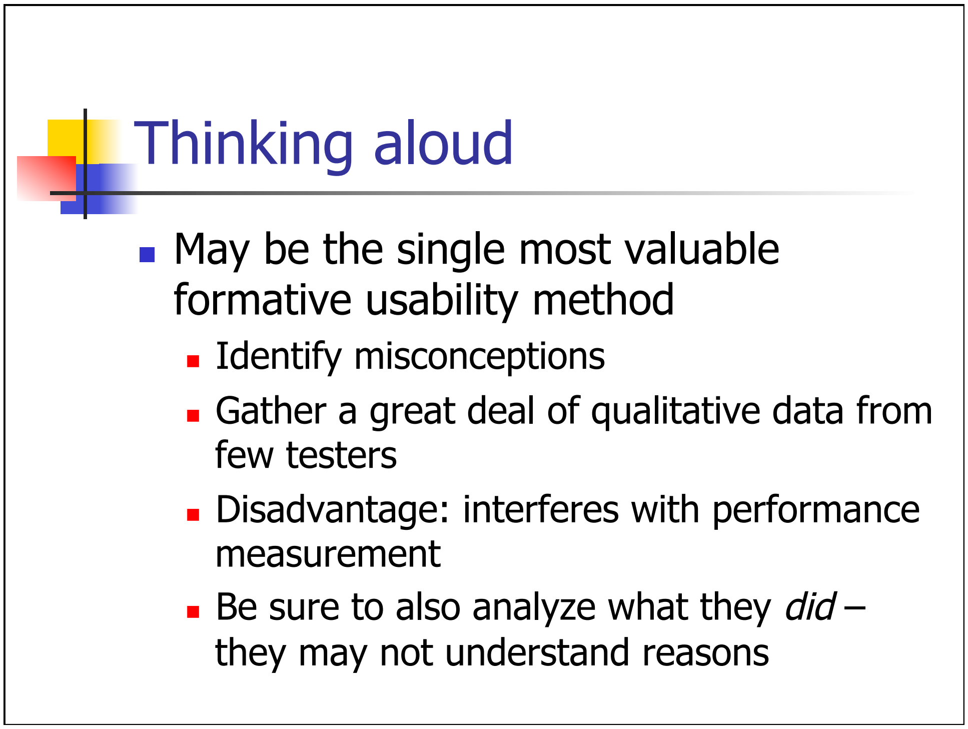
Data-Collection Technique:
Stopwatch
(with rules for when to start and stop the watch)

Be clear about the relationship between the specific measure you are using and the higher-level concept you are really interested in.



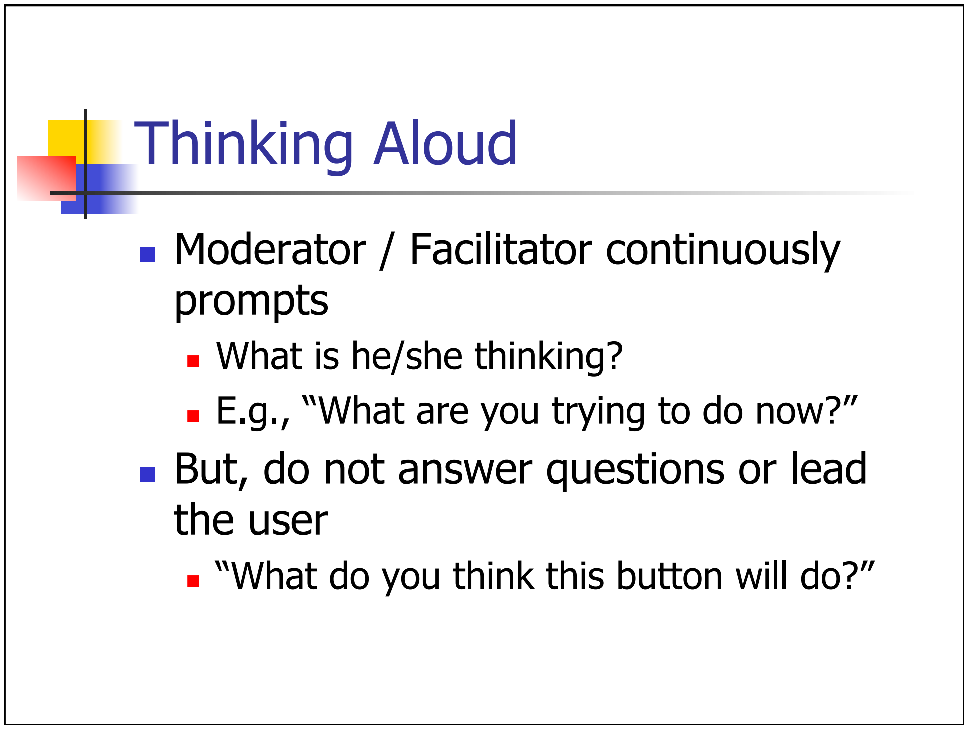
Performance Metrics

- Time to complete a task
- Number of tasks completed
- Time spent recovering from errors
- Number of errors
- Number of commands/functions used
  - Absolute or Unique
- Frequency of help use; time using
- Proportion who say they would use the product over a competitor’s
- Etc.



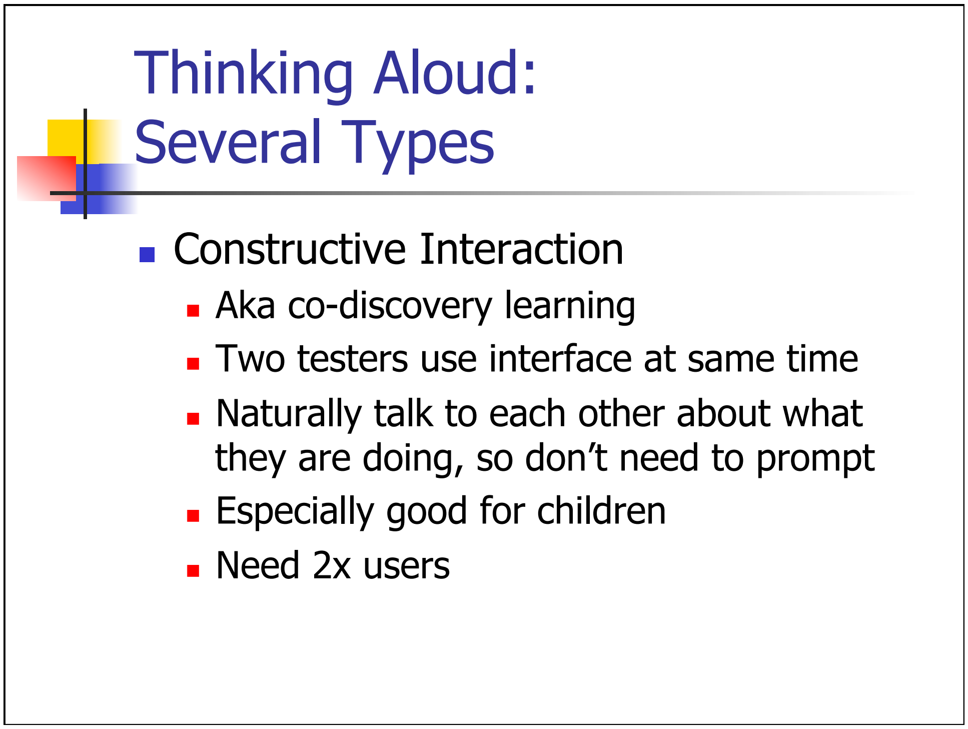
Thinking aloud

- May be the single most valuable formative usability method
  - Identify misconceptions
  - Gather a great deal of qualitative data from few testers
  - Disadvantage: interferes with performance measurement
  - Be sure to also analyze what they *did* – they may not understand reasons



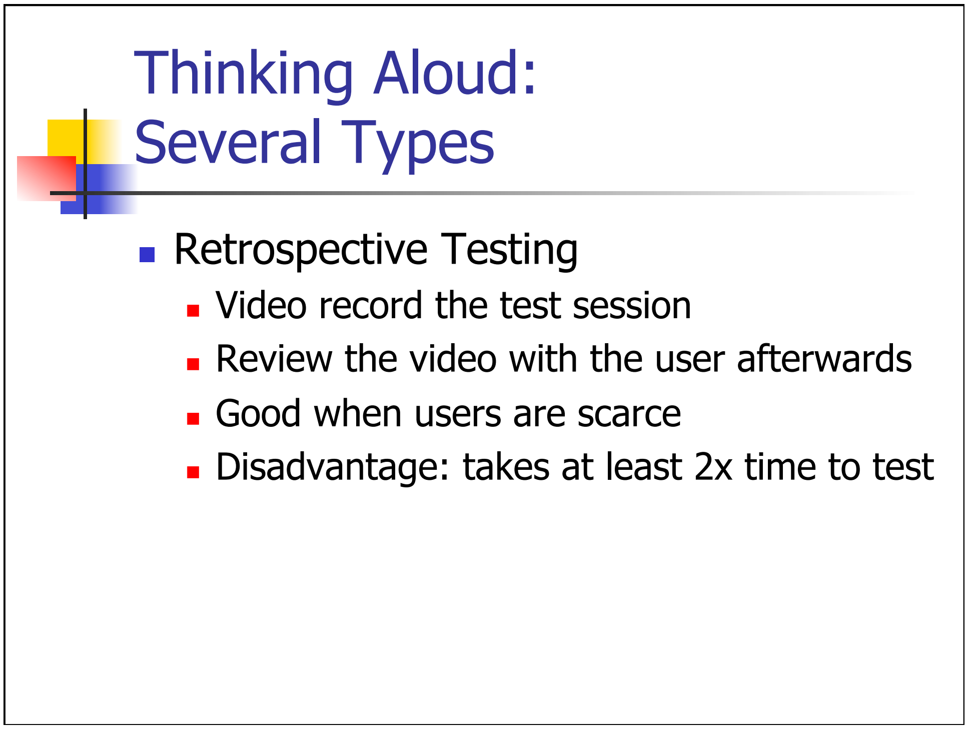
Thinking Aloud

- Moderator / Facilitator continuously prompts
  - What is he/she thinking?
  - E.g., "What are you trying to do now?"
- But, do not answer questions or lead the user
  - "What do you think this button will do?"



Thinking Aloud: Several Types

- Constructive Interaction
  - Aka co-discovery learning
  - Two testers use interface at same time
  - Naturally talk to each other about what they are doing, so don’t need to prompt
  - Especially good for children
  - Need 2x users

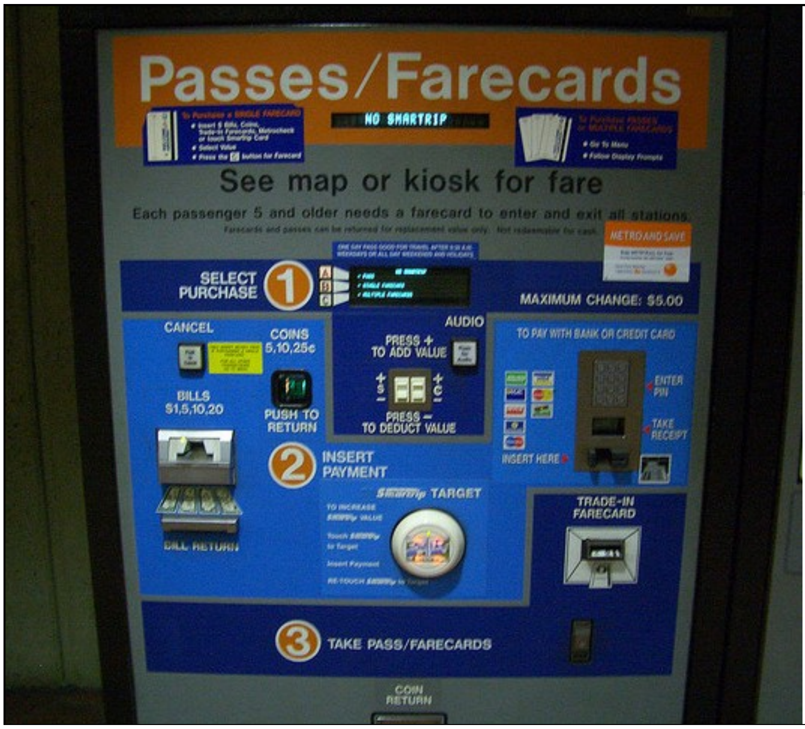


Thinking Aloud: Several Types

- Retrospective Testing
  - Video record the test session
  - Review the video with the user afterwards
  - Good when users are scarce
  - Disadvantage: takes at least 2x time to test

Thinking Aloud: Several Types

- Coaching
  - User can ask any questions of an “expert” coach.
  - Use to discover information needs of novice users
  - Use to develop training & help documentation



Which?

Exercise: Usability study of origami instructions

- Teams of 3+, 1 user, 1 moderator, N observers

To do

- Read
  - Ubicomp & Wearables (Benyon Ch 18 & 20).
- Start P8, P9
  - P8: usability test report, due 12/7
  - P9a: in-class oral presentation, 12/7
  - P9b: final report (cumulative), 12/12