Human-Computer Interaction IS4300

P7 – Heuristic Evaluation & Prototype Revision – Due

- After you receive the 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 – 2 weeks

- 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, 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

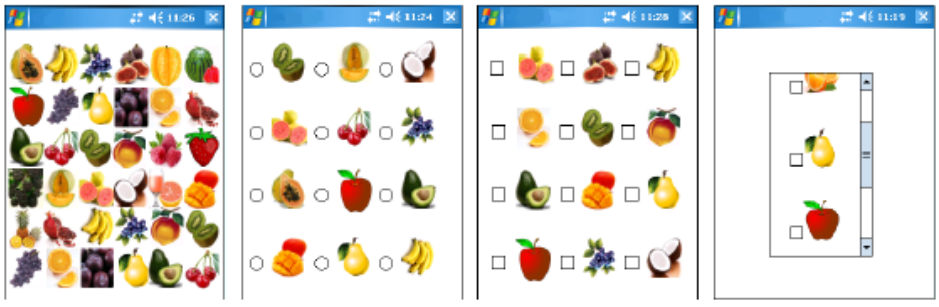
Research on Mobile UIs

Research Papers

- Mobile Interface Design for Low-Literacy Populations
- Multi-Layered Interfaces to Improve Older Adults' Initial Learnability of Mobile Applications

- Kind of study?
- Methodology?
- Main findings?

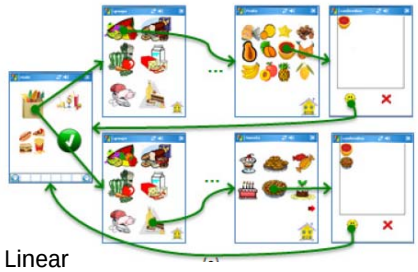
Mobile Interface Design for Low-Literacy Populations



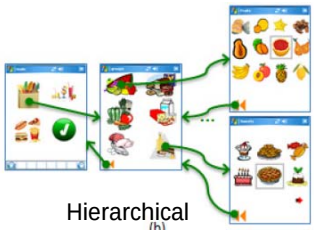
Study 1 – which widget is best?

- Icon vs. Radio Button vs. Checkbox vs. Scrollbar x 3 sizes
- N=17, all below 9th grade reading (REALM)
- Within subjects
- Results
  - Radio buttons best (performance & pref)
  - Large widgets best (performance & pref)

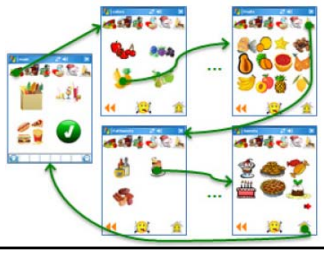
Study 2 – which navigation structure is best?



Linear



Hierarchical



Cross-linked

Study 2 – which navigation structure is best?

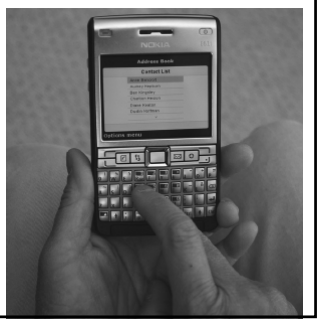
- N=19, low lit
- Users first trained on each interface
- Task = selecting a set of food items

- Results:
  - Linear is best (most tasks completed, most completed without error, recovered faster)
  - But – preferred cross-linked
  - Depth of 5, breadth 5-10 best (fewest errors)
  - Always provide BACK and HOME buttons

Multi-Layered Interfaces to Improve Older Adults' Initial Learnability of Mobile Applications

- "gray digital divide"
- Mobile devices require greater working memory (small UI, overloaded controls), which declines with age.
- Multi-Layered interface
  - "Training Wheels" aka scaffolding
  - Simplified interfaces decrease working memory load
  - May reduce abandonment of device

- N=16 older (65-81), 16 younger (21-36)
- Between subjects, stratified by age
  - ML: first master simple, then complex
  - Control: first master complex



Multi-Layered Interface Results

- ML simple could be learned in fewer steps
- ML simple resulted in better retention
- ML simple help elders more than younger users to master ML simple
- Elders rated ML simpler than control
- Elders preferred ML for learning simple tasks

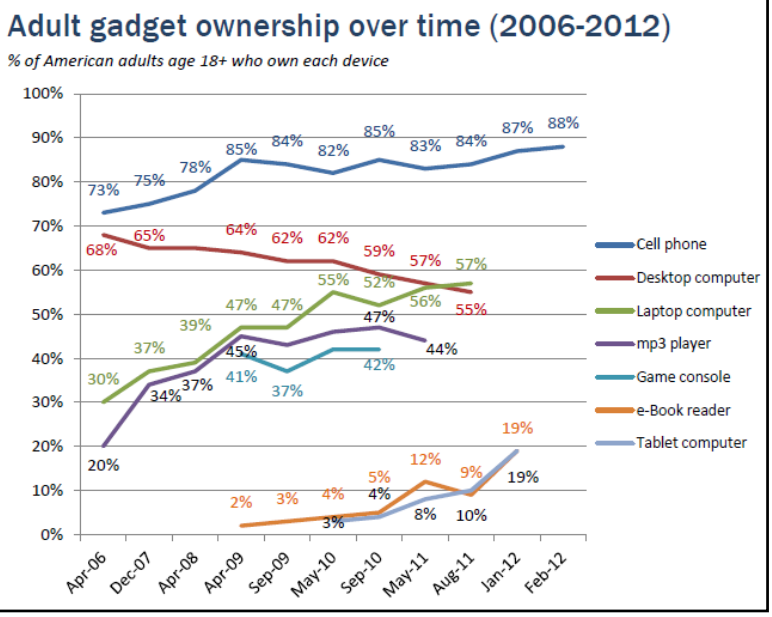
Summary

- Why are mobile interfaces for low literacy and elder users important?
- Are these two studies necessarily about mobile interfaces?

Digital Divide

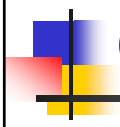
2012 (Pew)

SmartPhone penetration ~50% for most segments.



UI Evaluation Methods

- Expert/Inspection methods
  - Heuristic evaluation
  - Cognitive walk-through
  - Modeling
- User Testing
  - Qualitative methods (interviews, questionnaires, think aloud)
    - observation in the field
  - Quantitative methods
    - Descriptive studies
    - Experiments (same environment & task with 2 or more alternative designs)



Brief Review: Conducting Usability Studies

16

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

18

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

Crash course in human subjects research

23

Ethical Principles in Human Subjects Research (Belmont Report)

- Respect for persons (autonomy)
- Beneficience
- Justice

Northeastern University IRB

- Office of Research Regulatory Compliance
  www.research.neu.edu/research_integrity/
- Application process takes 1-2 months

IRB application not needed if...

- is a normal part of the students coursework;
- is supervised by a faculty member;
- has as its primary purpose the development of the student's research skills;
- does not present more than minimal risk to participants or to the student investigator;
- does not include any persons as research subjects under the age of 18;
- does not include any persons as research subjects who are classified as part of a vulnerable populations according to Federal regulations (see below);
- is not "genuine research" that is expected to result in publication or some other form of public dissemination;

You should obtain verbal consent – Example:

"Hi, we're designing a *XYZ*. *Explanation of XYZ.* We are conducting a study to find out what people think about this. We will not record or publish any information with your name. This is for a course we're taking in Human-Computer Interaction from Prof. Bickmore in the College of Computer and Information Science. Your participation is voluntary and you can stop anytime and ask that your data not be used. It should take about 30 minutes and we will compensate you with a can of Red Bull. Can you help us out with this?"

Nielsen on Usability Testing

Usability Engineering

Ch 6

Methodological Pitfalls

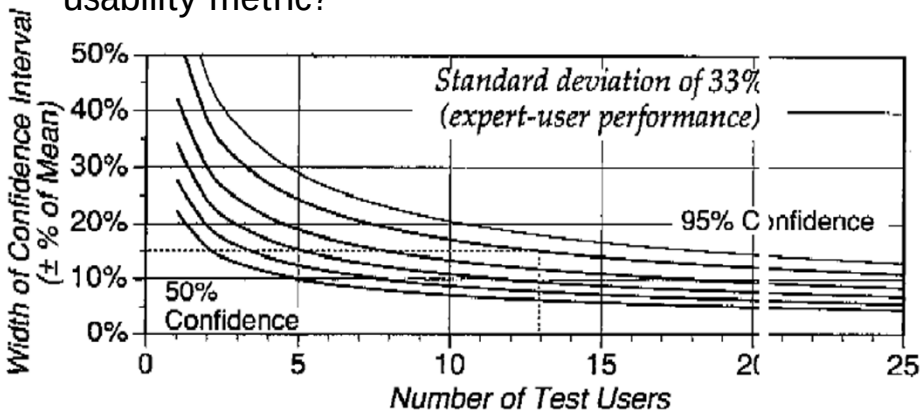
- Reliability
  - Test-retest
- Validity
  - Are the results correct and meaningful?

Reliability

- 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!

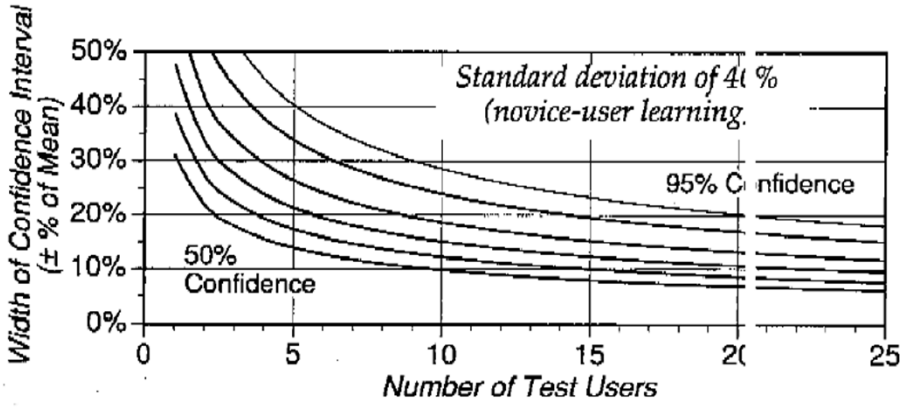
Another way to think about descriptive stats

- How many test users do I need to characterize a usability metric?



Another way to think about descriptive stats

- How many test users do I need to characterize a usability metric?



Sampling

Population

Mean?

Variance?

μ

σ^2

Sample of size N

$M = \frac{\sum X}{N}$

$SD^2 = \frac{\sum (X - M)^2}{N}$

Mean values from all possible samples of size N

aka "distribution of means"

μ

NOTE: This is a normal curve

$\sigma_M^2 = \frac{\sigma^2}{N}$

33

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?)

Type of Errors in Hypothesis Testing

	"The Truth"	
	H0 True	**H0 False**
Decide to Reject H0	**Type I Error**	**Correct Decision**
Do not Reject H0	**Correct Decision**	**Type II Error**

'p' = Probability of Type I Error

36

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.

37

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

38

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

39

Sampling Techniques

- *Systematic Sampling*
  - Every k^{th} 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

40

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

41

Advanced Sampling Techniques

- *Cluster Sampling*
  - Used when populations are very large
  - The unit of sampling is a group (e.g., a class in a school) rather than individuals
  - Groups are randomly sampled from the population (e.g., ten classes from a particular school)

42

Advanced Sampling Techniques

- *Multistage Sampling*
  - Variant of cluster sampling
  - First, identify large clusters (e.g., school districts) and randomly sample from that population
  - Second, sample individuals from randomly selected clusters
  - Can be used along with stratified sampling to ensure a representative sample

43

Sampling

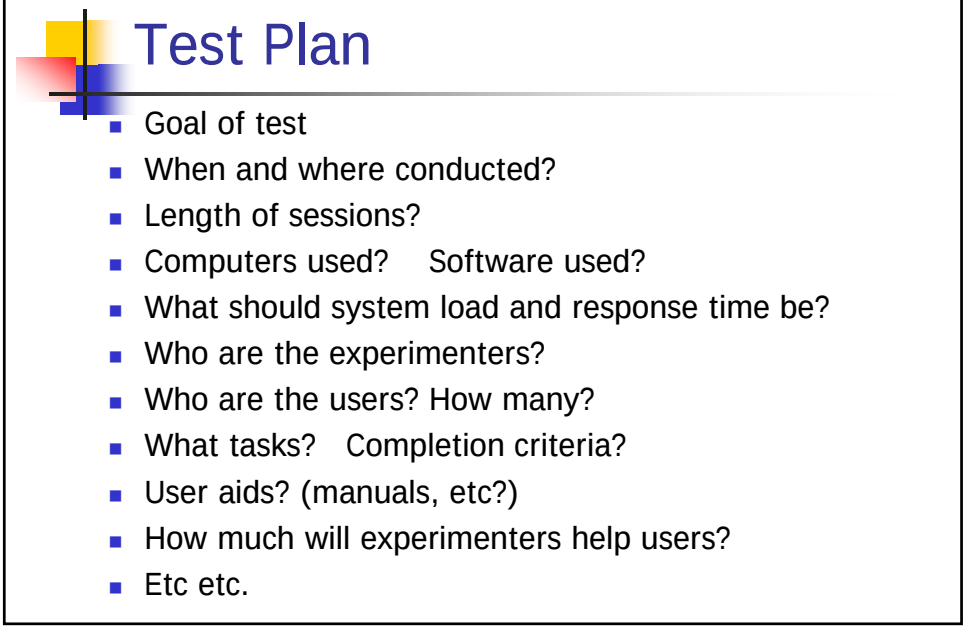
- Most statistics assume a random sample.

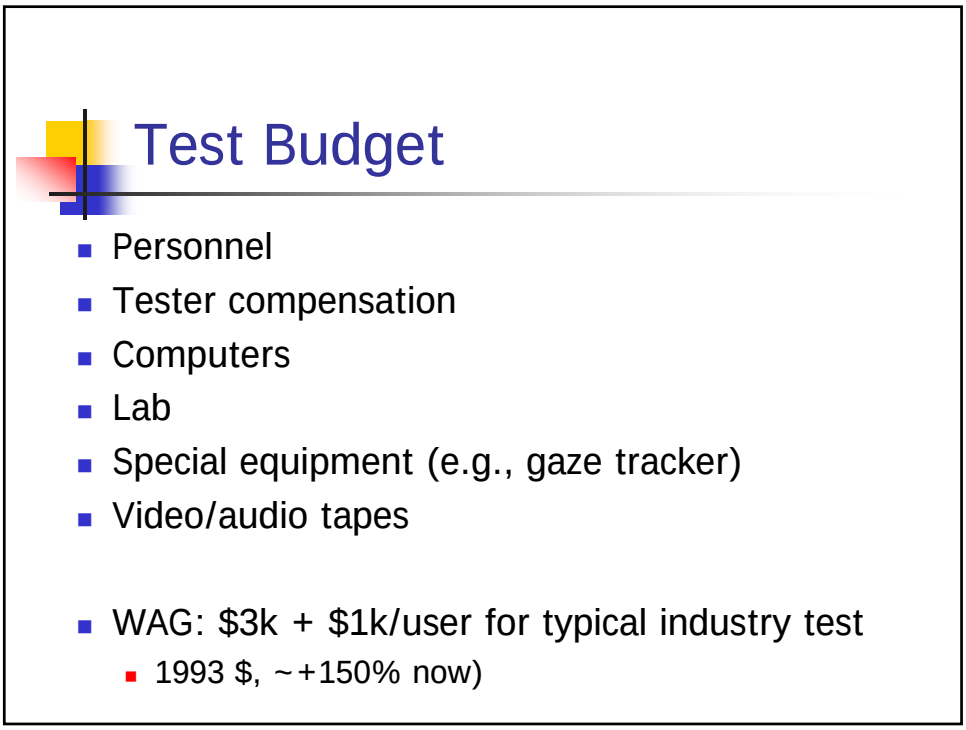
44

Sample size

- In all empirical research, you should motivate your *sample size*
- Formative usability testing:
  - 3-5 test users => 80% of bugs
- Summative Experimental testing:
  - Do a statistic power analysis
  - Google “power analysis calculator”

45

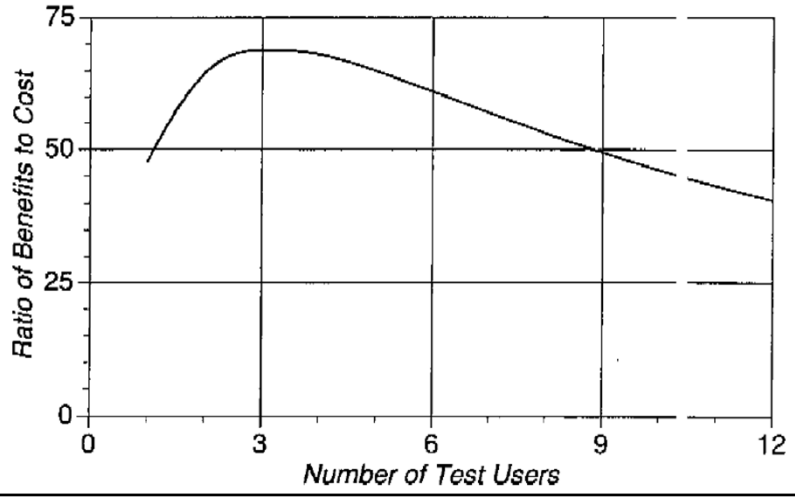

Test Plan
Goal of test
When and where conducted?
Length of sessions?
Computers used? Software used?
What should system load and response time be?
Who are the experimenters?
Who are the users? How many?
What tasks? Completion criteria?
User aids? (manuals, etc?)
How much will experimenters help users?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


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 an given user)
- Examples
  - Value of finding a usability problem = $15k
  - N = 41
  - $\lambda = 0.31$

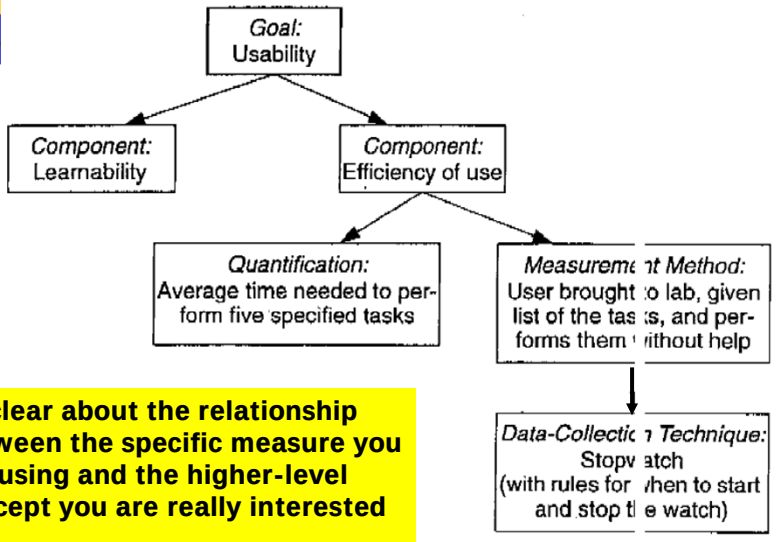
Payoff ratio given these assumptions



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



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 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

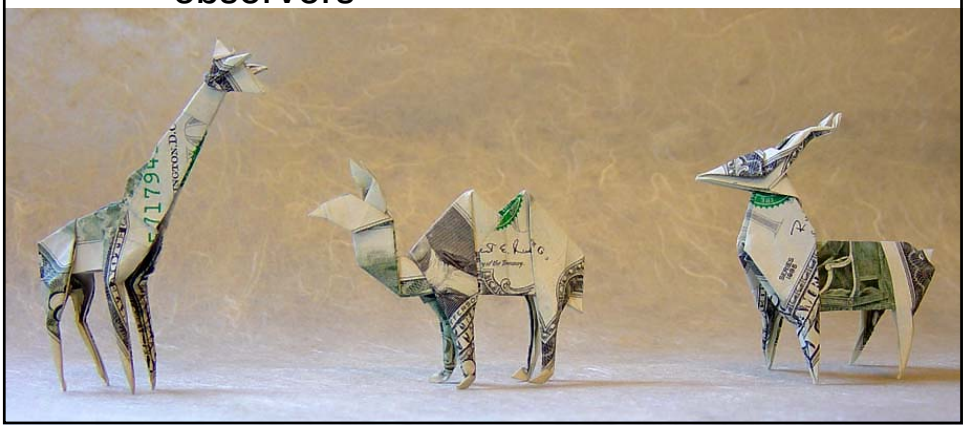

Passes/Farecards
NO SMARTRIP
See map or kiosk for fare
Each passenger 5 and older needs a farecard to enter and exit all stations.
METRO AND SAVE
SELECT PURCHASE
1
MAXIMUM CHANGE: $5.00
CANCEL
COINS 5,10,25¢
AUDIO
PRESS + TO ADD VALUE
PRESS – TO DEDUCT VALUE
TO PAY WITH BANK OR CREDIT CARD
ENTER PIN
TAKE RECEIPT
INSERT HERE
BILLS $1,5,10,20
PUSH TO RETURN
2
INSERT PAYMENT
SmarTrip TARGET
TRADE-IN FARECARD
BILL RETURN
3
TAKE PASS/FARECARDS
COIN RETURN

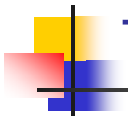

Passes/Farecard
Phone

Which?

Exercise: Usability study of origami instructions

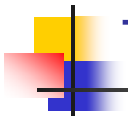
- Teams of 3+, 1 user, 1 moderator, N observers





P8 – Finish Project & Do User Testing – 2 weeks

- 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, 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



To do

- Read
  - Designing for the Web
    - Dix Ch 21
  - Start P8, P9