Usability Engineering

Chapter 6
Usability Testing

Jakob Nielsen
Morgan Kaufmann, 1993
ISBN 978-0-12-518406-9

Introduction

- Reliability
  - Questions of whether one would get the same result if the test were to be repeated
- Validity
  - Questions of whether the result actually reflects the usability one wants to test

Reliability

- Possible to have huge individual differences between tests
  - Best user is 10x faster than slowest user
  - Best 25% of users are 2x faster than slowest 25% of users

User A, Interface X
Task Time – 3 min 18 sec

User B, Interface Y
Task Time – 5 min 30 sec

Interface X is 40% faster, does this mean interface X is better?

How do we know that User A isn't just generally faster than User B?

If we repeated the test with User C and User D, will the results be the same?

3

Reliability

- Often need to make decision on the basis of fairly unreliable data
  - Some data is better than no data
- Must choose between Interface X and Interface Y
  - May choose Interface X, some evidence in it's favor
  - Number of statistical methods available

4

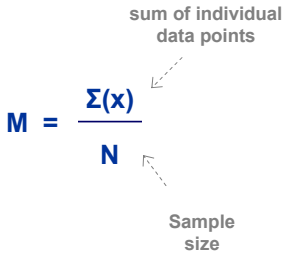
Reliability

- Survey of 36 usability studies
  - 33% MSD for measures of expert-user performance
  - 46% MSD for measures of novice-user learning
  - 59% MSD for error rates
- What do we care?
  - Helps to define number of users needed to achieve the desired confidence interval

5

Reliability

- Mean
  - Sum of the observations divided by the number of observations
  - Identifies central location of the data



Find the mean of the following data set: 3,2,4,1,4,4

$$M = \frac{3 + 2 + 4 + 1 + 4 + 4}{6} = \frac{18}{6} = 3$$

6

Reliability

- Mean Standard Deviation
  - Most common measure of variability, measuring the spread of the data set and the relationship of the mean to the rest of the data
  - Small SD, fairly uniform data
  - Large SD, wide variance in data

individual data point — mean of all scores

$$S^2 = \frac{\Sigma(X-M)^2}{N-1}$$

Sample size

Find the Standard deviation of the following data set: 3,2,4,1,4,4

X	M	(X-M)	$(X-M)^2$
3	3	0	0
2	3	-1	1
4	3	1	1
1	3	-2	4
4	3	1	1
4	3	1	1

$$S^2 = \frac{0 + 1 + 1 + 4 + 1 + 1}{6-1}$$

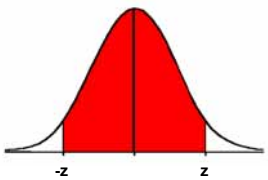
$$S^2 = \frac{8}{5} = 1.6$$

$$S = 1.265$$

7

Reliability

- Confidence Interval
  - Estimated range of values which is likely to include an unknown population parameter
  - Estimated range being calculated from a given set of sample data
  - Provides some idea about how uncertain we are about the unknown parameter



Confidence Level = 95%

Confidence Interval = [-z, z]

Fraction of Data	Number of Standard Deviations from Mean
50.0%	0.674
68.3%	1.000
90.0%	1.645
95.0%	1.960
95.4%	2.000
98.0%	2.326
99.0%	2.576
99.7%	3.000

8

Reliability

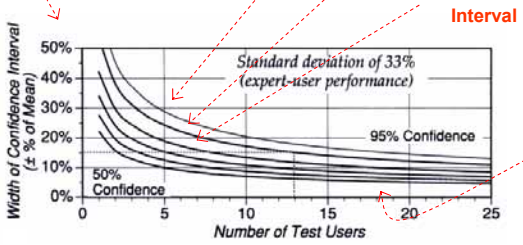
The confidence interval is plus or minus that many percent of the measure mean value of the usability metric

95% Confidence Interval

90% Confidence Interval

80% Confidence Interval

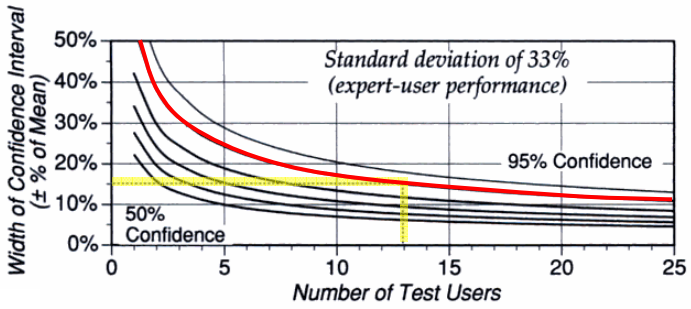
50% Confidence Interval



- Confidence level of 95% typically used for research studies, assuming normal distribution

9

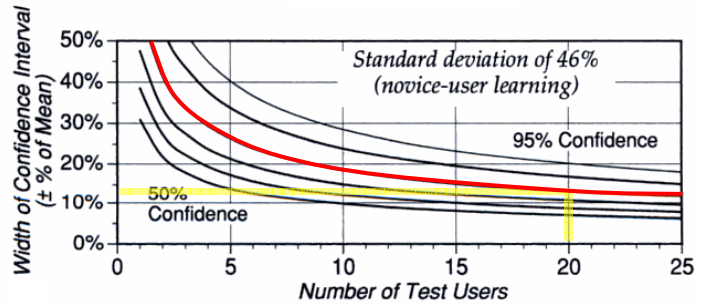
Reliability



- Measure expert-user performance
  - Want 90% chance that true value is no more than 15% different from the mean value measured in the usability test, how many users are needed?
  - Start with 15% on the y-axis, find the corresponding point using the 90% confidence level curve
  - Need about 13 users

10

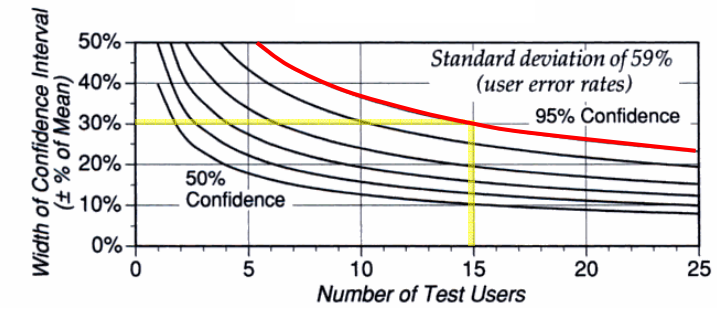
Reliability



- Measure novice-user learning
  - Want 80% chance that true value is no more than 12% different from the mean value measured in the usability test, how many users are needed?
  - Start with 12% on the y-axis, find the corresponding point using the 80% confidence level curve
  - Need about 20 users

11

Reliability

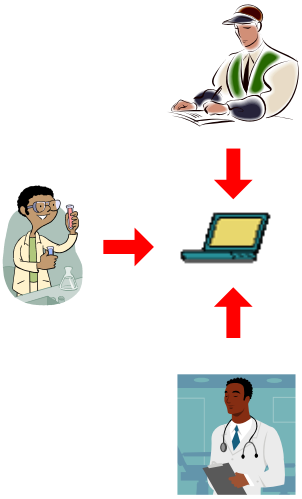


- Measure user error rates
  - Want 95% chance that true value is no more than 30% different from the mean value measured in the usability test, how many users are needed?
  - Start with 15% on the y-axis, find the corresponding point using the 90% confidence level curve
  - Need about 15 users

12

Validity

- Does usability test measure something of relevance?
  - Requires methodological understanding of the test method used
- Typical methods
  - Using wrong users
  - Giving users wrong tasks



13

Validity

- Confounding effects
  - Moving from a character based interface to a graphical interface
  - Test using a 24x80 alphanumeric screen verses a 1024x1024 pixel graphics display
  - Results reflect comparison between the screens as well as comparison between character and graphical user interfaces

14

Performance Measurement

- Major pitfall – measuring something that is poorly related to the property of interest
  - Simplified model relates true goal of measurement study to data collection
- Usability is abstract concept
  - Break down into usability attributes
- Once components identified, how to measure them?
  - Quantify efficiency in terms of how long it takes to perform five tasks
- Once we quantify, how do we measure?
  - Bring users into lab and give them a list of tasks to perform
  - Observe users in normal environment and measure whenever a task like the on specified occurs

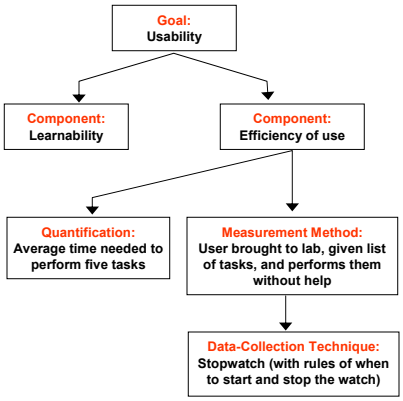


Figure 19 – Model of usability measurement

15

Performance Measurement

- Typical quantifiable usability measures

- Time user takes to complete a specific task
- Number of tasks (or the proportion of a larger task) that can be completed within a given time limit
- Ratio of successful interactions and errors
- Time spent recovering from errors
- Number of user errors
- Number of immediately subsequent erroneous actions
- Number of command or other features utilized by the user (absolute or number of different commands)
- Number of system features user can remember during a debriefing after the test
- Frequency of the use of manual and/or help system, and time spent on these elements
- Frequency the manual and/or help system solved the user's problem
- Proportion of user statements during the test that were positive versus critical toward the system
- Number of times the user expresses clear frustration (or joy)
- Proportion of users who say that they would prefer using the system over some specified competitor
- Number of times the user had to work around an unsolvable problem
- Proportion of user using efficient working strategies compared to users who use inefficient strategies (when multiple choices available)
- Amount of "dead" time when user is not interaction with the system (response time, or thinking time)
- Number of times user is sidetracked form focusing on the real task

16

Thinking Aloud

- Having test subject use the system while continuously thinking aloud
  - Verbalizing thoughts enable designers to understand how they view the system
  - May give false impression if too much weight is given to users' own "theories", still important to make notes of what users are doing
  - Method is used to what users are doing and whey they are doing it while they are doing it to avoid later rationalizations
- Potential to impact results
  - Need to verbalize can slow users down
  - Problem solving behavior influenced by the fact they are verbalizing thoughts

User overlooks a certain field in the dialog box

User says, "I would have seen it immediately if it were higher in the dialog box."

Observer notices that user wasn't looking in the dialog box

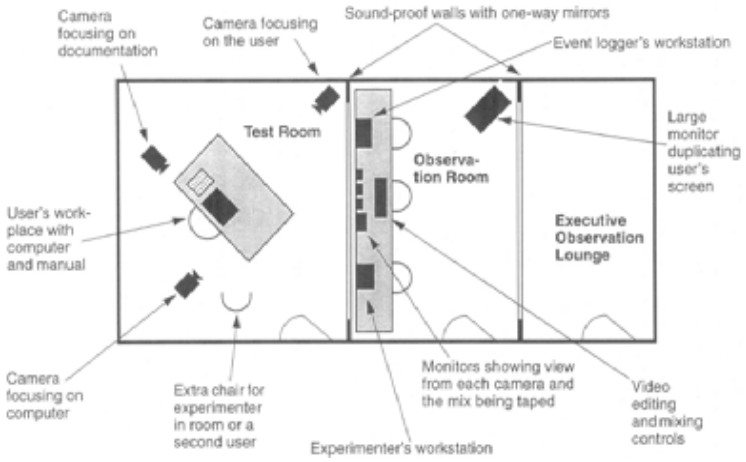
17

Variations of the Thinking Aloud Method

- Constructive Interaction (codiscovery learning)
  - Two test users use a system together
  - More natural test situation
  - Users may make more comments when trying to solve problem
  - Drawback – users with different style that cannot work together
- Retrospective Testing
  - Record test session, have user review the recording
  - More extensive comments from user
  - Facilitator can stop take and ask more detailed questions without interfering with test
- Coaching Method
  - Facilitator tries to steer the user in the right direction
  - User is allowed to ask any system related question, expert coach will answer to best of their ability
  - Able to discover the information needs of a user to provide better training and documentation

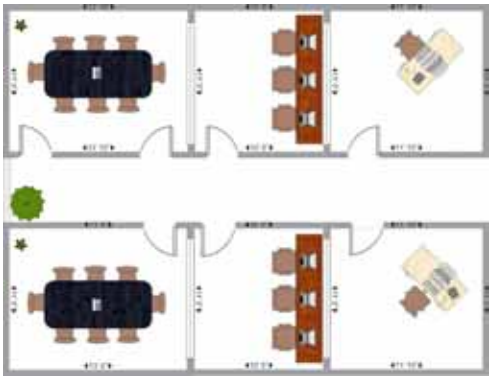
18

Usability Laboratories



19

Usability Laboratories



20

Usability Laboratories

- Portable Usability Laboratories
  - Allows for more flexible testing and field studies
  - Don't need a dedicated space
- True Discount Portable Laboratory
  - Notepad to take notes
  - Laptop to run software being tested
- Hallway Method
  - User interface placed on display in a heavily trafficked area
  - Suggest test task to user
  - Record task time and/or comments

21

Additional Topics

- Some of the topics in this chapter have already been covered, we'll skip these in lecture
  - Test Goals and Test Plans
  - Getting Test Users
  - Choosing Experimenters
  - Ethical Aspects of Test with Human Subjects
  - Test Tasks
  - Stages of a Test
- You should still read them over

22