Introduction

• Learn about usability of product, ask user

• Many ways to ask these questions
  • Rating scale
  • Attribute list
  • Open-ended questions
  • Many others…

• Bottom line – you are asking participant for information
  • Self-reported metrics

Subjective data

Counterpart to objective data
Subjective to each participant providing input, however objective from perspective of usability specialist

Preference data

Counterpart to performance data
Not necessarily a choice of one option over another
Importance of Self-Reported Data

- Provides information about users’ perception of the system and their interaction with it
- Main goal can be how users feel about the system
  - Can take forever to perform task
  - Experience makes them happy
- Example – travel planning
  - Want user to think of your product first
  - Not likely to remember how many clicks to finish task
  - Remember they liked it last time, more likely to return

Collecting Self-Reported Data

- Likert Scales
  - Expresses degree of agreement with statement
  - Uses an odd number of response options, allowing neutral response

<table>
<thead>
<tr>
<th>AGREEMENT</th>
<th>FREQUENCY</th>
<th>IMPORTANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly Agree</td>
<td>Always</td>
<td>Very Important</td>
</tr>
<tr>
<td>Agree</td>
<td>Very Frequently</td>
<td>Important</td>
</tr>
<tr>
<td>Undecided</td>
<td>Occasionally</td>
<td>Moderately Important</td>
</tr>
<tr>
<td>Disagree</td>
<td>Rarely</td>
<td>Of Little Importance</td>
</tr>
<tr>
<td>Strongly Disagree</td>
<td>Very Rarely</td>
<td>Unimportant</td>
</tr>
<tr>
<td></td>
<td>Never</td>
<td></td>
</tr>
</tbody>
</table>

- Semantic Differential Scales
  - Presenting pairs of bipolar, opposite adjectives

<table>
<thead>
<tr>
<th>QUALITY</th>
<th>LIKELIHOOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely Poor</td>
<td>Almost Always True</td>
</tr>
<tr>
<td>Below Average</td>
<td>Usually True</td>
</tr>
<tr>
<td>Average</td>
<td>Often True</td>
</tr>
<tr>
<td>Above Average</td>
<td>Occasionally True</td>
</tr>
<tr>
<td>Excellent</td>
<td>Sometimes But Infrequently True</td>
</tr>
<tr>
<td></td>
<td>Usually Not True</td>
</tr>
<tr>
<td></td>
<td>Almost Never True</td>
</tr>
</tbody>
</table>
Collecting Self-Reported Data

- When to collect
  - End of each task
  - End of the entire session

- Post-task ratings
  - Quick ratings help to pinpoint problematic parts of the interface

- Post-study ratings
  - More common
  - Allows for in-depth ratings and open-ended questions
  - Effective overall evaluation after interacting with product more fully

How to collect

- Answer questions or provide ratings orally
- Record responses on a paper form
- Provide responses using an online tool

Biases in Collecting Self-Reported Data

- More positive when asked in person than anonymous web survey

Social desirability bias

- Responders give answers they believe will make them look better in the eyes of others or not disappoint the evaluator

Collect data

- In a way that moderator or facilitator does not see responses until after participant has left
- Anonymous
Collecting Self-Reported Data

- General Guidelines for Rating Scales
  - Assess specific attribute such as visual appeal, credibility, or responsiveness
  - Data more reliable if you ask participants to assess attribute in different ways
    - Average scores together to get overall reaction
    - Questionnaires have both positive and negative statements
  - Number of scale values – odd versus even
    - Odd allows for neutral response
    - Even forces user towards one end of scale

- Analyzing Self-Reported Data
  - Assign numeric value to each position on the scale
    - Compute averages
  - Top-2-boxes and Bottom-2-boxes
    - Not interval data, should report as frequency

Can we do this? Do we have interval data?

Not technically, most researchers treat it as interval data

Assume distance between 1 & 2 is the same distance between 2 & 3

Degrees of intervalness
Post-Task Ratings

- Associate ratings with each task
  - Provide insight into which tasks were most difficult

- Ease of Use
  - Most common self-reported metric
  - How easy/difficult each task was

Please indicate your opinion about completing tasks

Easy  ○ ○ ○ ○ ○  Difficult
• • •

Post-Task Ratings

- After-Scenario Questionnaire (ASQ)
  - Jim Lewis (1991) developed set of three rating scales
  - Each statement accompanied by 7-point rating scale
  - Each question touches on three fundamental areas of usability

<table>
<thead>
<tr>
<th>Effectiveness?</th>
<th>1. Overall, I am satisfied with the ease of completing the tasks in this scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficiency?</td>
<td>2. Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario</td>
</tr>
<tr>
<td>Satisfaction?</td>
<td>3. Overall, I am satisfied with the support information (on-line help, messages, documentation) when completing the tasks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly disagree</td>
<td>○ ○ ○ ○ ○ ○ ○</td>
<td>Strongly agree</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neither Agree or Disagree</td>
<td>○ ○ ○ ○ ○ ○ ○</td>
<td>Neither Agree or Disagree</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strongly Agree</td>
<td>○ ○ ○ ○ ○ ○ ○</td>
<td>Strongly Agree</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Post-Task Ratings

- **Expectation Measure**
  - More important is how easy/difficult task was in comparison to easy/difficult user thought it would be

- **Process**
  - User rate how easy/difficult they expected each tasks to be
  - Perform each task
  - User rate how easy/difficult task actually was
  - Display two scores on scatter plot

**Expectation rating – “before” rating**

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Difficult</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Very Easy</td>
</tr>
</tbody>
</table>

**Experience rating – “after” rating**

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Difficult</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Very Easy</td>
</tr>
</tbody>
</table>

Post-Task Ratings

- **Four Quadrants**

  - **Lower right**
    - Thought – easy / Actual – difficult
    - Biggest dissatisfiers or disappointments
    - Tasks you should focus on first

  - **Upper right**
    - Thought – easy / Actual – easy
    - Product works fine, don’t break it!

  - **Upper left**
    - Thought – difficult / Actual – easy
    - Pleasant surprises for users and designers
    - Could be “features” of system

  - **Lower left**
    - Thought – difficult / Actual – difficult
    - No surprises
    - Big opportunities to make improvements
Post-Task Ratings

- **Magnitude Estimation**
  - McGe (2004) argued departure from traditional rating-scale approach and use estimation from psychophysics
  - User provided with reference (e.g. light source) and assign value or attribute
  - User provided with new stimuli (e.g. brighter light source) and assign value in comparison to reference
  - **KEY** – User maintains a correspondence between ratios of numbers they assign and their perception of the magnitude

![Reference 10x brighter than reference](image1.jpg)

![Reference 30x brighter than reference](image2.jpg)

Post-Task Ratings

- **Usability Magnitude Estimation**
  - Give "good design" and "bad design" for same task
  - Assign "usability values" to good and bad reference design (any positive number)
  - As task performed make judgments compared to reference values

- **Modification**
  - No reference designs
  - User starts giving usability values for tasks as they encounter them
  - Gradually build own ruler as they progress

- **Presentation**
  - Data transformed using geometric averaging (log transformations) to a consistent scale for comparison purposes
Post-Task Ratings

- How do the different Post-Task Self-Reported Metrics compare?

Condition 1: “Overall, this task was: Very Difficult ○ ○ ○ ○ Very Easy”

Condition 2: “Please rate the usability of the site for this task: Very Difficult to Use ○ ○ ○ ○ Very Easy to Use”

Condition 3: “Overall, I am satisfied with the ease of completing this task: Strongly Disagree ○ ○ ○ ○ Strongly Agree”

Condition 4 (before doing all tasks): “How difficult or easy do you expect this task to be? Very Difficult ○ ○ ○ ○ Very Easy”

Condition 5: “How difficult or easy did you find this task to be? Very Difficult ○ ○ ○ ○ Very Easy”

Condition 5 (after doing each task): “How difficult or easy did you find this task to be? Very Difficult ○ ○ ○ ○ Very Easy”

Condition 5: “Please assign a number between 1 and 100 to represent how well the website supported you for this task. Remember, 1 would mean that the site was not at all supportive and completely unusable. A score of 100 would mean that the site was perfect and would require absolutely no improvement.”

Post-Task Ratings

- Method
  - Participants performed 6 tasks on a live application to lookup information about employees
  - Each participant used only one of the five self-report techniques
  - Total of 1,131 employees participated with at least 210 using each self-report technique

- Collected task time and binary success (Fig 6.4)
  - Task 2 most challenging
  - Task 4 easiest

- Average task ratings across all conditions (Fig 6.5)
  - Significant correlation between performance data and ratings

Figure 6.4 – Performance data.

Figure 6.5 – Average subjective ratings across all techniques. Ratings are expressed as a percentage of the maximum possible rating.
Post-Task Ratings

- Averages of task rating for each of the tasks, split by condition (Fig 6.6)
  - Pattern similar regardless of the technique used
  - At large sample sizes all five techniques can effectively distinguish between tasks
- Correlation between sample size (Fig 6.7)
  - Random samples of different sizes
  - Condition 1 results in better correlation starting at the smallest sample size and continuing
  - Simplest scale most reliable at smaller sample sizes

![Figure 6.6 – Average subjective ratings split by task and condition](image1)

![Figure 6.7 – Subsampling analysis showing average correlation between ratings for the six tasks from subsamples of various sizes and the full dataset for each condition](image2)

Post-Session Ratings

- Overall measure of perceived usability
  - Self-reported metrics collected after having completed their interaction with the product
- Take average of self-reported data across all tasks
  - Looking at how perception changes over time
- Collect data once at end of session
  - Measuring participant’s last impression of experience
    - Likely to influence future decision about product
Post-Session Ratings

- System Usability Scale
  - Developed by Brooke in 1986
  - Consists of 10 statements users rate their level of agreement
  - Half positively worded
  - Half negatively worded

- Calculating SUS score
  - Each item contributes a score of 0 to 4
    - Items 1, 3, 5, 7, and 9
      - Score = scale position – 1
    - Items 2, 4, 6, 8, and 10
      - Score = 5 - scale position
  - Sum values and multiply by 2.5

<table>
<thead>
<tr>
<th>Item</th>
<th>Scale Position</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>

Total = 22
SUS Score = 22 x 2.5 = 55

Post-Session Ratings

- Computer System Usability Questionnaire (CSUQ)
  - Developed by Lewis (1995) who also developed ASQ technique for post-task ratings
  - Overall assessment of a system at the end of a usability study
  - 19 questions to which user rates agreement on a 7-point scale
  - All statements are worded positively
  - Results viewed in four main categories
    - System Usefulness
    - Information Quality
    - Interface Quality
    - Overall Satisfaction
Post-Session Ratings

Questionnaire for User Interface Satisfaction
- Developed by Human-Computer Interaction Laboratory at University of Maryland (Chin, Diehl, & Norman, 1998)

- Five categories
  - Overall Reaction
  - Screen
  - Terminology/System Information
  - Learning
  - System Capabilities

- 27 Questions on a 10-point scale, anchors change with statement
  - First six have no statements, polar opposites to rate overall reaction
Post-Session Ratings

- **Usefulness, Satisfaction, and Ease of Use Questionnaire**
  - Lund (2001) proposed USE questionnaire
  - 30 rating scales divided into four categories
  - All positive statements
  - Users respond with level of agreement on 7-point Likert scale
Post-Session Ratings

- Product Reaction Cards
  - Benedek and Miner (2002)
  - Presented a set of 118 cards containing adjectives
  - User chooses cards they felt described the system
  - Narrow down to top five, and explain why they choose each
  - Purpose is to elicit commentary

- Comparison of Post-Session Self-Reported Metrics
  - Tullis and Stetson (2004) compared variety of post-session questionnaires for measuring user reactions to websites in an online usability study
  - Two web portals evaluated
  - Total of 123 participants
  - Each participant used one of the questionnaires to evaluate both websites

SUS - Word system in every question replaced with website
QUIS - scales not appropriate to website dropped. Word system replaced with website and term screen replaced with webpage
CSUQ - Term system or computer replaced by website
Product Reaction Cards - Each word presented with checkbox, choose as many as they wished
Study Questionnaire - Nine positive statements, respond on 7 point Likert scale
Post-Session Ratings

- Comparison of Post-Session Self-Reported Metrics (continued)

Authors analyzed all data and found site 1 got significantly better ratings than site 2.

What would results be at sample size 6 to 14?
- Size = 6, only 30-40% identified site 1 is preferred.
- Size = 8 (common size), SUS identified site 1 as preferred 75% of the time.

SUS yield more consistent ratings at relatively sample sizes.

May be from use of both positive and negative statements.

Doesn’t try to break down assessment into more detailed components, just assessment of site as a whole.

What Is A Good or Bad SUS Score?

- Review of a large number of published usability studies.
  - Over 50 studies reported average SUS scores across 129 different conditions.
  - Covered wide range of subjects:
    - Websites, applications, computer hardware, mobile devices, voice systems.
  - Conducted in various parts of the world:
    - US, Germany, Switzerland, UK, New Zealand.
  - Users per study ranged from 5 – 81.

- Distribution of scores shown (right):
  - Average = 66%
  - Median = 69%

- Results:
  - Below 60% is relatively poor.
  - Above 80% is pretty good.
Visualizing Data Using Radar Charts

- Some techniques for capturing self-reported data yield values on several dimensions

- Assume you have data from USE questionnaire
  - Usefulness = 90%
  - Satisfaction = 50%
  - Ease of Use = 40%
  - Ease of Learning = 40%

- Plot these values as radar chart
  - Choose “Radar” as the “Chart Type” in Excel
  - Help viewer easily detect patterns represented by different shapes

Online Services

- More companies appreciate getting feedback from users
  - Voice of the Customer

- Same process as in post-session self-reported metrics
  - Difference is evaluation done on a live website
  - Randomly selected user offered a pop-up survey
  - Provide standard mechanism for getting feedback at various places on site
Online Services

- Website Analysis and Measurement Inventory (WAMMI)
  - Designed to evaluate websites (versus software applications or multi-media systems)
  - 20 statements with associated 5-point Likert scale
  - Both positive and negative statements
  - Scale already tested on hundreds of websites, you can compare to reference database

<table>
<thead>
<tr>
<th>Strongly disagree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>This website has much that is of interest to me.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>It is difficult to move around this website.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>I can quickly find what I want on this website.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>This website seems logical to me.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>This website needs more introductory explanations.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>The pages on this website are very attractive.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>I feel in control when I'm using this website.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>This website is too slow.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>This website helps me find what I am looking for.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>Learning to find my way around this website is a problem.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>I don't like using this website.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>I can easily contact the people I want to on this website.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>I feel efficient when I'm using this website.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>It is difficult to tell if this website has what I want.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>Using this website for the first time is easy.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>The website has some annoying features.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>Remembering where I am on this website is difficult.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>Using this website is a waste of time.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>I get what I expect when I click on things on this website.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
<tr>
<td>Everything on this website is easy to understand.</td>
<td>○ ○ ○ ○ ○</td>
</tr>
</tbody>
</table>

Results divided into five areas
- Attractiveness
- Controllability
- Efficiency
- Helpfulness
- Learnability
- Plus and overall usability score

Each score is standardized from reference to their database
- Average score = 50
- Perfect score = 100
### Online Services

- **American Customer Satisfaction Index (ACSI)**
  - Developed at Stephen M. Ross Business School of The University of Michigan
  - Covers variety of industries including automotive, retail, manufacturing

- **ForeSee applies ACSI to online experience and produce industry-specific indices**
  - Popular to analyzing US government websites
  - Annual Top 40 Online Retail Satisfaction assess Amazon.com, Netflix, L.L. Bean, J. C. Penney, Dell, CompUSA

- **Methodology**
  - Core set of 14 to 20 questions customized to function of website
  - Asks for a rating on a 10-point scale of different attributes
  - Additional custom questions added to profile site visitors in terms of visit intent, visit frequency, and other specific information that helps organization profile visitors in terms meaningful to their business or mission

### Online Services Form

<table>
<thead>
<tr>
<th>Question</th>
<th>Rating Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Please rate the quality of information on this site.</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>2. Please rate the freshness of content on this site.</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>3. Please rate the convenience of the services on this site.</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>4. Please rate the ability to accomplish what you wanted to on this site.</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>5. Please rate the clarity of site organization.</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>6. Please rate the clean layout of this site.</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>7. Please rate the ability to find information you want on this site.</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>8. Please rate the clarity of site map / directory</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>9. Please rate the reliability of site performance on this site.</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>10. What is your overall satisfaction with this site?</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>11. How does this site meet your expectations?</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>12. How does this site compare to your idea of an ideal website?</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>13. How likely are you to return to this site?</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>14. How likely are you to recommend this site to someone else?</td>
<td>1 2 3 4 5 6 7 8 9 10 Don't Know</td>
</tr>
<tr>
<td>15. How frequently do you visit this site?</td>
<td>Please Select</td>
</tr>
<tr>
<td>16. What would you like to see improved on our site? (optional)</td>
<td></td>
</tr>
</tbody>
</table>
Online Services

- American Customer Satisfaction Index (continued)
  - Results divided into six quality-related elements that drive satisfaction
    - Content
    - Functionality
    - Look & Feel
    - Navigation
    - Search
    - Site Performance
  - Show relative impact of improving satisfaction for each element on increasing satisfaction overall
  - Future behavior scores
    - Likelihood to return
    - Recommend to others
  - All scores on a 100-point index scale
    - Assessment of the impact of each score shown in four quadrant chart
    - Scores in lower right (high impact, low score) indicate where to focus

Online Services

- OpinionLab
  - Provides page-level feedback from users
  - Page-level floating icon stays in the bottom right corner
    - Click on icon leads user to one of the methods shown in Fig 6.23

Figure 6.22 – Animated icon stays in bottom right corner of the page regardless of scroll position

Figure 6.23(a) – Allows user to give page a quick overall rating

Figure 6.23(b) – Allows user to give more detailed feedback on a few different scales
Online Services

- OpinionLab (continued)
  - Variety of ways to visualize data
  - Shows the most-rated 200 pages graphically
  - Most negative ratings on left, neutral ratings at top, most positive ratings on right

- Issues with Live-Site Surveys
  - Number of Questions
    - Fewer the questions, the higher response rate
    - Ask if every question is “must have” information
    - Some researchers believe 20 is the maximum number of questions you should ask in this type of survey
  - Self-selection of respondents
    - Respondents get to decide about whether they get to complete the survey
    - Should ask if this biases the responses in any way
    - Unhappy people more likely to respond than happy/satisfied users
    - If goal is to uncover problems, then unhappy people are not a problem
  - Number of respondents
    - Many services work on the basis of percentage of visitors to decide when to offer survey
    - Monitor responses closely to see if percentage need to change
  - Nonduplication of respondents
    - Most services provide mechanism for noting when survey already offered
    - Prevents duplication of responses from individual and not annoying users who don’t want to respond
Other Types of Self-Reported Metrics

- Self-reporting techniques so far assess user reactions to products/websites as a whole or to tasks performed using them.

- Study objective may want to assess user reactions to specific product attributes overall or specific product elements.

- Assessing Specific Attributes
  - Visual appeal
  - Perceived efficiency
  - Usefulness
  - Enjoyment
  - Credibility
  - Appropriateness of terminology
  - Ease of navigation
  - Responsiveness

Figure 6.25 – Flash website images for 50 to 500 ms and rated on visual appeal. Results consistent across times.

Figure 6.26 – Artificially manipulated load times, user rate on 5-point scale of unacceptable/acceptable.

Other Types of Self-Reported Metrics

- Assessing Specific Elements
  - Instructions
  - FAQs
  - Online help
  - Homepage
  - Search Function
  - Site Map

- Assessing subjective reactions to specific elements same as for specific aspects.

- Tullis (1998) focused on candidate homepage designs
  - Participants shown templates and rated on 5-point scale (-2, -1, 0, 1, 2)

Figure 6.26 – Artificially manipulated load times, user rate on 5-point scale of unacceptable/acceptable.
Other Types of Self-Reported Metrics

- Open-ended questions
  - Most usability studies include some open-ended questions in addition to rating scales
  - Common technique is to allow user to add comments relating to any of the individual rating scales
  - Ask participant to list 3-5 things they liked the most about the product and 3-5 things they liked the least about the product
    - Can be translated into metrics by counting instances items listed

- Blur distinction between self-reported data and performance data
  - Awareness
    - User interacts with product/website
    - Questionnaire giving list of specific pieces of content site may/may not have
    - Asked which content user saw or remembered
    - Users not allowed to refer back
  - Comprehension
    - User interacts with product/website
    - Given quiz to test comprehension of some information on site
    - If possible knowledge pre-existing, important to administer pre-test
Other Types of Self-Reported Metrics

- Awareness and Usefulness Gaps
  - Look at difference between participants’ awareness of a specific piece of information or functionality and perceived usefulness of the same piece of information or functionality.
  - Majority of users unaware of certain functionality, but once they notice it they find it useful.
    - This functionality should be promoted.

- Measuring
  - Need both awareness and usefulness metric.
  - Ask user about awareness as a yes/no question.
  - Give users a couple of minutes to explore functionality.
  - Ask on a scale of 1-5 how useful is this functionality.
  - Plot percentage of participants who are aware of the functionality next to percent who found function useful (top-2 box).

Summary

- Getting self-reported data both at a task level and at the end of the usability session.
  - Task-level data can help identify areas that need improvement.
  - Session-level data can help get a sense of overall usability.

- When testing in a lab.
  - Consider using one of the standard questionnaires for assessing subjective reactions.
  - System Usability Scale shown robust even with small numbers of participants (e.g., 8-10).

- When testing a live website.
  - Consider using online services for measuring user satisfaction.
  - Major advantage is the ability to compare to their reference database.

- Be creative in the use of other techniques in addition to simple rating scales.
  - Ask for ratings on a given topic in several different ways and average the results to get more consistent data.
  - Carefully construct any new rating scales.
  - Make appropriate use of open-ended questions.
  - Consider techniques like checking for awareness or comprehension after interaction with the product.