

Measuring the User Experience

Collecting, Analyzing, and Presenting Usability Metrics

Chapter 3 Planning a Usability Study

Tom Tullis and Bill Albert

Morgan Kaufmann, 2008
ISBN 978-0123735584

Introduction

- What are the goals of your usability study?
 - Are you trying to ensure optimal usability for a new piece of functionality?
 - Are you benchmarking the user experience for an existing product?
- What are the goals of users?
 - Do users complete a task and then stop using the product?
 - Do users use the product numerous times on a daily basis
- What is the appropriate evaluation method?
 - How many participants are needed to get reliable feedback?
 - How will collecting metric impact the timeline and budget?
 - How will the data be collected and analyzed?

Study Goals

- How will the data be used within the product development lifecycle?
- Two general ways to use data
 - Formative
 - Summative



Formative

Chef who periodically checks a dish while it's being prepared and makes adjustments to positively impact the end result



Summative

Evaluating the dish after it is completed like a restaurant critic who compares the meal with other restaurants

3

Study Goals

- Formative Usability
 - Evaluates product or design, identifies shortcomings, makes recommendations
 - Repeats process
- Attributes
 - Iterative nature of testing with the goal of improving the design
 - Done before the design has been finalized
- Key Questions
 - What are the most significant usability issues that are preventing users from completing their goals or that are resulting in inefficiencies?
 - What aspects of the product work well for users? What do they find frustrating?
 - What are the most common errors or mistakes users are making?
 - Are improvements being made from one design iteration to the next?
 - What usability issues can you expect for remain after the product is launched?



Formative

Chef who periodically checks a dish while it's being prepared and makes adjustments to positively impact the end result

4

Study Goals

- **Summative Usability**
 - Goal is to evaluate how well a product or piece of functionality meets its objectives
 - Comparing several products to each other
 - Focus on evaluating against a certain set of criteria
- **Key Questions**
 - Did we meet the usability goals of the project?
 - How does our product compare against the competition?
 - Have we made improvements from one product release to the next?



Summative

Evaluating the dish after it is completed like a restaurant critic who compares the meal with other restaurants

5

User Goals

- **Need to know about users and what they are trying to accomplish**
 - Forced to use product everyday as part of their jobs?
 - Likely to use product only one or twice?
 - Is product a source of entertainment?
 - Does user care about design aesthetic?
- **Simplifies to two main aspects of the user experience**
 - Performance
 - Satisfaction



6

User Goals

- Performance
 - What the user does in interacting with the product
- Metrics (more in Ch 4)
 - Degree of success in accomplishing a task or set of tasks
 - Time to perform each task
 - Amount of effort to perform task
 - Number of mouse clicks
 - Cognitive effort
- Important in products that users don't have choice in how they are used
 - If user can't successfully complete key tasks, it will fail



7

User Goals

- Satisfaction
 - What users say or think about their interaction
- Metrics (more in Ch 6)
 - Ease of use
 - Exceed expectations
 - Visually appealing
 - Trustworthy
- Important in products that users have choice in usage



8

Choosing the Right Metrics

Ten Types of Usability Studies

- Every usability study has unique qualities, ten scenarios provided with recommendations for each

Usability Study Scenario	Task Success	Task Time	Errors	Efficiency	Learnability	Issue-Based Metrics	Self-Reported Metrics	Behavioral and Physiological Metrics	Combined and Comparative Metrics	Live Website Metrics	Card-Sorting Data
1. Completing a transaction	X			X		X	X			X	
2. Comparing products	X			X			X		X		
3. Evaluating frequent use of the same product	X	X		X	X		X				
4. Evaluating/navigation and/or information architecture	X		X	X							X
5. Increasing awareness							X	X		X	
6. Problem discovery						X	X				
7. Maximizing usability for a critical product	X		X	X							
8. Creating an overall positive user experience							X	X			
9. Evaluating the impact of subtle changes											X
10. Comparing alternative designs	X	X				X	X		X		

9

Choosing the Right Metrics

Ten Types of Usability Studies

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

Issue Based Metrics (Ch 5)

- Anything that prevents task completion
- Anything that takes someone off course
- Anything that creates some level of confusion
- Anything that produces an error
- Not seeing something that should be noticed
- Assuming something should be correct when it is not
- Assuming a task is complete when it is not
- Performing the wrong action
- Misinterpreting some piece of content
- Not understanding the navigation

10

Choosing the Right Metrics

Ten Types of Usability Studies

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

Self Reported Metrics (Ch 6)

- Asking participant for information about their perception of the system and their interaction with it
 - Overall interaction
 - Ease of use
 - Effectiveness of navigation
 - Awareness of certain features
 - Clarity of terminology
 - Visual appeal
 - Likert scales
 - Semantic differential scales
 - After-scenario questionnaire
 - Expectation measures
 - Usability Magnitude Estimation
 - SUS
 - CUSQ (Computer System Usability Scale)
 - QUIS (Questionnaire for User Interface Satisfaction)
 - WAMMI (Website Analysis & Measurement Inventory)
 - Product Reaction Cards

11

Choosing the Right Metrics

Ten Types of Usability Studies

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

Behavioral and Physiological Metrics (Ch 7)

- Verbal Behaviors
 - Strongly positive comment
 - Strongly negative comment
 - Suggestion for improvement
 - Question
 - Variation from expectation
 - Stated confusion/frustration
- Nonverbal Behaviors
 - Frowning/Grimacing/Unhappy
 - Smiling/Laughing/Happy
 - Surprised/Unexpected
 - Furrowed brow/Concentration
 - Evidence of impatience
 - Leaning in close to screen
 - Fidgeting in chair
 - Rubbing head/eyes/neck

• • •

12

Choosing the Right Metrics

Ten Types of Usability Studies

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

Combined and Comparative Metrics (Ch 8)

- Taking smaller pieces of raw data like task completion rates, time-on-task, self reported ease of use to derive new metrics such as an overall usability metric or usability score card
- Comparing existing usability data to expert or idea results

13

Choosing the Right Metrics

Ten Types of Usability Studies

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

Live Website Metrics (Ch 9)

- Information you can glean from live data on a production website
 - Server logs – page views and visits
 - Click through rates - # times link shown vs. actually clicked
 - Drop off rates – abandoned process
 - A/B studies – manipulate the pages users see and compare metrics between them

Card Sorting Data (Ch 9)

- Open card sort
 - Give participants cards, they sort and define groups
- Closed card sort
 - Give participants cards and name of groups, they put cards into groups

14

Choosing the Right Metrics

Ten Types of Usability Studies

- Completing a Transaction
 - Make transaction as smooth as possible
 - Well-defined beginning and end
 - Completing purchase, registering new software, selling stocks
- Metrics
 - Task Success
 - Scored as success or failure
 - Need clear end-state to confirm task success
 - Reporting percentage of success good measure of effectiveness of transaction
 - Efficiency
 - User must complete same transaction many time, good to measure task completion per unit of time
 - Issue Based Metrics
 - Issue severity assigned to each usability issue, able to identify and focus on high-priority problems
 - Self Reported Metrics
 - Likelihood to return
 - User expectations
 - Live Website Metrics (if transaction involves a website)
 - Drop-off rate from transaction may indicate problematic steps in transaction

Metrics

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

15

Choosing the Right Metrics

Ten Types of Usability Studies

- Comparing Products
 - How does product compare to the competition?
 - How does product compare to previous releases?
- Metrics
 - Task Success
 - Being able to complete a task correctly is essential
 - Efficiency
 - Gauged with task completion time, number of page views, number of action steps taken
 - How much effort is required to use the product
 - Self Reported Metrics
 - Provide a good summary of user's overall experience
 - Satisfaction makes sense when user has choice of products
 - Combined and Comparative Metrics
 - Want a big picture of how product compares from usability perspective

Metrics

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

16

Choosing the Right Metrics

Ten Types of Usability Studies

- Evaluating Frequent Use of the Same Product
 - Need to be easy and highly efficient
 - Microwave, DVD players, web applications
 - Most people have little patience for products that are difficult to use

- Metrics
 - Task Success
 - Task Time
 - Measuring time to complete set of core tasks
 - Reveal effort involved
 - Helpful to compare task completion time to expert
 - Efficiency
 - Number of steps need
 - Time may be short, but separate decision can be numerous
 - Learnability
 - Time/effort required to achieve maximum efficiency
 - Measure is previous efficiency metrics over time
 - Self Reported Metrics
 - Awareness and usefulness
 - Find aspects of product that should be highlighted
 - Low awareness, but once they find it, they find it's extremely useful

Metrics

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

17

Choosing the Right Metrics

Ten Types of Usability Studies

- Evaluating Navigation and/or Information Architecture
 - Users can quickly and easily find what they are looking for
 - Navigate around the product
 - Know where they are within the overall structure
 - Know what options are available to them

- Metrics
 - Task Success
 - Typically done very early in design with wire frames or partially functioning prototypes
 - Give participants key pieces of information to see how well interface works
 - Errors
 - Efficiency
 - Lostness – number of steps it took to complete a task relative to minimum number
 - Card Sorting Data
 - Understand how participants organize information
 - Closed sort identifies how many items placed into correct category and indicates intuitiveness of information architecture

Metrics

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

18

Choosing the Right Metrics

Ten Types of Usability Studies

- Increasing Awareness
 - Aimed at increasing awareness of a specific piece of content or functionality
 - Why is something not noticed or used?
- Metrics
 - Live Website Metrics
 - Monitor interactions
 - Not foolproof – user may notice and decide not to click, alternatively user may click but not notice interaction
 - A/B testing to see how small changes impact user behavior
 - Self Reported Metrics
 - Pointing out specific elements to user and asking whether they had noticed those elements during task
 - Aware of feature before study began
 - Not everyone has good memory
 - Show users different elements and ask them to choose which one they saw during task
 - Behavioral and Physiological Metrics
 - Eye tracking
 - Determine amount of time looking at a certain element
 - Average time spent looking at a certain element
 - Average time before user first noticed it

Metrics

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

19

Choosing the Right Metrics

Ten Types of Usability Studies

- Problem Discovery
 - Identify major usability issues
 - After deployment, find out what annoys users
 - Periodic checkup to see how users are interaction with the product
- Discovery vs. usability study
 - Open-ended
 - Participants may generate own tasks
 - Strive for realism in typical task and in user's environment
 - Comparing across participants can be difficult
- Metrics
 - Issue Based Metrics
 - Capture all usability issues, you can convert into type and frequency
 - Assign severity rating and develop a quick-hit list of design improvements
 - Self Reported Metrics

Metrics

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

20

Choosing the Right Metrics

Ten Types of Usability Studies

- Maximizing Usability for a Critical Product
 - Instead of *striving* to be easy to use and efficient (cell phone), some product *have* to be (defibrillator, emergency exit instructions on airplane)
- Critical vs. noncritical product
 - Entire reason for product's existence is for user to complete a very important task
 - Failure will have a significant negative outcome
 - Important that user performance measured against a target goal
 - If it doesn't meet goal, it must be redesigned
 - Need larger number of users to have high degree of certainty
- Metrics
 - Errors
 - Number of errors or mistakes
 - Important to be explicit of what an error is and isn't
 - Task Success
 - Binary approach recommended
 - Efficiency
 - May also want to tie success to other metrics such as competition time without errors
 - Defibrillator example...

Metrics

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

21

Choosing the Right Metrics

Ten Types of Usability Studies

- Creating an Overall Positive User Experience
 - Not enough to be usable, want exceptional user experience
 - Thought provoking, entertaining, slightly-addictive
 - Performance useful, but what user thinks, feels, and says really matters
- Metrics
 - Self Reported
 - Satisfaction – common but not enough
 - Exceed expectations – want user to say it was easier, more efficient, or more entertaining than expected
 - Likelihood to purchase, use in future
 - Recommend to a friend
 - Behavioral and Physiological
 - Pupil diameter
 - Heart rate
 - Skin conductance

Metrics

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

22

Choosing the Right Metrics

Ten Types of Usability Studies

- Evaluating the Impact of Subtle Changes
 - Impact on user behavior may not be clear, but huge implications to the larger population
 - Font choice and size, placement, visual contrast, color, image choices
 - Terminology, content
- Metrics
 - Live Website Metrics
 - A/B tests compares control design against alternative
 - Compare traffic or purchases to the controlled design

Metrics

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

23

Choosing the Right Metrics

Ten Types of Usability Studies

- Comparing Designs
 - Comparing more than one design alternative
 - Early in the design process teams put together semi-functional prototypes
 - Evaluate using predefined set of metrics
- Participants
 - Can't ask same participant to perform same tasks with all designs
 - Even with counterbalancing design and task order, information on valuable
- Procedure
 - Study as between-subjects, participant only works with one design
 - Have primary design participant works with, show alternative designs and ask for preference

Metrics

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

24

Choosing the Right Metrics

Ten Types of Usability Studies

- Comparing Designs (continued)
- Metrics
 - Task Success
 - Indicates which design more usable
 - Small sample size, limited value
 - Task Time
 - Indicates which design more usable
 - Small sample size, limited value
 - Issue Based Metrics
 - Compare the frequency of high-, medium-, and low-severity issues across designs to see which one most usable
 - Self Reported Metrics
 - Ask participant to choose the prototype they would most like to use in the future (forced comparison)
 - As participant to rate each prototype along dimensions such as ease of use and visual appeal

Metrics

- Task Success
- Task Time
- Errors
- Efficiency
- Learnability
- Issue Based Metrics
- Self Reported Metrics
- Behavioral and Physiological Metrics
- Combined and Comparative Metrics
- Live Website Metrics
- Card Sorting Data

25

Other Study Details

- Budgets and Timelines
 - Difficult to provide cost or time estimates for a any particular type of study
- General rules of thumb
 - Formative study
 - Small number of participants (≤ 10)
 - Little impact
 - Lab setting with larger number of participants (> 12)
 - Most significant cost – recruiting and compensating participants
 - Time required to run tests
 - Additional cost for usability specialists
 - Time to clean up and analyze data
 - Online study
 - Half of the time is spent setting up the study
 - Running online study requires little if any time for usability specialist
 - Other half of time spent cleaning up and analyzing data
 - 100-200 person-hours (50% variation)



26

Other Study Details

- Evaluation Methods
 - Not restricted to certain type of method (lab test vs. online test)
 - Choosing method based on how many participants and what metrics you want to use
- Lab test with small number of participants
 - One-on-one session between moderator and participant
 - Participant thinking-aloud, moderator notes participant behavior and responses to questions
 - Metrics to collect
 - Issue based metrics – issue frequency, type, severity
 - Performance metrics – task success, errors, efficient
 - Self-reported metrics – answer questions regarding each task at the end of study
 - Caution
 - Easy to over generalize performance and self-reported metrics without adequate sample size



27

Other Study Details

- Evaluation Methods (continued)
- Lab test with larger number of participants
 - Able to collect wider range of data because increased sample size means increased confidence in data
 - All performance, self-reported, and physiological metrics are fair game
 - Caution
 - Inferring website traffic patterns from usability lab data is not very reliable
 - Looking at how subtle design changes impact user experience
- Online studies
 - Testing with many participants at the same time
 - Excellent way to collect a lot of data in a short time
 - Able to collect many performance, self reported metrics, subtle design changes
 - Caution
 - Difficult to collect issue-based data, can't directly observe participants
 - Good for software or website testing, difficult to test consumer electronics



28

Other Study Details

- Participants
 - Have major impact in findings
- Recruiting issues
 - Identifying the recruiting criteria to determine if participant eligible for study
 - How to segment users
 - How many users are needed
 - Diversity of user population
 - Complexity of product
 - Specific goals of study
 - Recruiting strategy
 - Generate list from customer data
 - Send requests via email distribution lists
 - Third party
 - Posting announcement on website



29

Other Study Details

- Data Collection
 - Plan how you are capturing data needed for study
 - Significant impact on how much work later when analysis begins
- Lab test with small number of participants
 - Excel works well
 - Have template in place for quickly capturing data during testing
 - Data entered in numeric format as much as possible
 - 1 – success
 - 0 – failure
 - Everyone should know coding scheme extremely well
 - Someone flips scales or doesn't understand what to enter
 - Throw out data or have to recode data
- Larger studies
 - Use data capture tool
 - Helpful to have option to download raw data into excel



30

Other Study Details

- Data Cleanup
 - Rarely in a format that is instantly ready to analyze
 - Can take anywhere from one hour to a couple of weeks
- Cleanup tasks
 - Filtering data
 - Check for extreme values (task completion times)
 - Some participants leave in the middle of study, and times are unusually large
 - Impossible short times may indicate user not truly engaged in study
 - Results from users who are not in target population
 - Creating new variables
 - Building on raw data useful
 - May create a top-2-box variable for self-reported scales
 - Aggregate overall success average representing all tasks
 - Create an overall usability score



31

Other Study Details

- Cleanup tasks (continued)
 - Verifying responses
 - Notice large percentage of participants giving the same wrong answer
 - Check why this happens
 - Checking consistency
 - Make sure data capture properly
 - Check task completion times and success to self reported metrics (completed fast but low rating)
 - Data captured incorrectly
 - Participant confused the scales of the question
 - Transferring data
 - Capture and clean up data in Excel, then use another program to run statistics, then move to Excel to create charts and graphs



32

Summary

- Formative vs. summative approach
 - Formative – collecting data to help improve design before it is launched or released
 - Summative – want to measure the extent to which certain target goals were achieved
- Deciding on the most appropriate metrics, take into account two main aspects of user experiences – performance and satisfaction
 - Performance metrics – characterize what the user does
 - Satisfaction metrics - relate to what users think or feel about their experience
- Budgets and timelines need to be planned well out in advance when running any usability study
- Three general types of evaluation methods used to collect usability data
 - Lab tests with small number of participants
 - Best for formative testing
 - Lab test with large number of participants (>12)
 - Best for capturing a combination of qualitative and quantitative data
 - Online studies with very large number of participants (>100)
 - Best to examine subtle design changes and preferences

33

Summary

- Clearly identify criteria for recruiting participants
 - Truly representative of target group
 - Formative
 - 6 to 8 users for each iteration is enough
 - If distinct groups, helpful to have four from each group
 - Summative
 - 50 to 100 representative users
- Plan how you are going to capture all the data needed
 - Template for quickly capturing data during test
 - Everyone familiar with coding conventions
- Data cleanup
 - Manipulating data in a way to make them usable and reliable
 - Filtering removes extreme values or records that are problematic
 - Consistency checks and verifying responses make sure participant intentions map to their responses

34