Measuring the User Experience

Collecting, Analyzing, and Presenting Usability Metrics

Chapter 4
Performance Metrics

Tom Tullis and Bill Albert

Morgan Kaufmann, 2008
ISBN 978-0123735584

Introduction

- Anyone who uses technology has to interact with some type of interface to accomplish their goals
  - The way users behave or interact with a product forms the cornerstone of performance metrics
  - Every type of user behavior is measureable in some way
- Performance metrics rely on user behaviors and the use of tasks or scenarios
- Useful to estimate the magnitude of a specific usability issue
  - Not enough to know there is an issue, but how many people are likely to encounter this issue
- Able to tell ***what*** was effective (or not), but not ***why***
- Five basic types
  - Task success, Time-on-task, Errors, Efficiency, Learnability

2

Task Success

- Measures how effectively users able able to complete a given set of tasks
- Provide as binary success and levels of success
- Almost universal metric because it can be calculated for a wide variety of things being tested
- Easy to relate to

3

Task Success

- Collecting any Type of Success Metric
  - Must have clear end state
  - Need to know what constitutes a success
- How to collect
  - Have user verbally articulate the answer after completing the task
    - May provide extra or arbitrary information
  - Provide answer using online tool or paper form
    - Avoid write-in answers
    - Time consuming to analyze each answer, may involve judgment calls
  - Provide multiple choice responses
  - Proxy measure
    - Response depends on individual users, not there in person to verify
    - Ask participant to write title of page that shows balance

Task with clear end state

Find the current price for a share of Google Stock

Task with no so clear end state

Research ways to save for your retirement

4

Task Success

- Binary Success
  - Simplest and most common way to measuring task success
  - Participants complete task or they don't

	A	B	C	D	E	F
1	Participant	Task1	Task2	Task3	Task4	Task5
2	P1	1	0	1	0	0
3	P2	1	0	1	0	1
4	P3	1	1	1	1	1
5	P4	1	1	1	1	1
6	P5	0	0	1	1	1
7	P6	1	0	0	1	1
8	P7	0	1	1	1	1
9	P8	0	0	1	1	0
10	P9	1	0	1	0	1
11	P10	1	1	1	1	1
12	P11	0	1	1	1	1
13	P12	1	0	1	1	1
14	Average	67%	42%	92%	75%	83%
15	Confidence Interval (95 %)	28%	22%	29%	29%	29%

0 = Task failure

1 = Task success

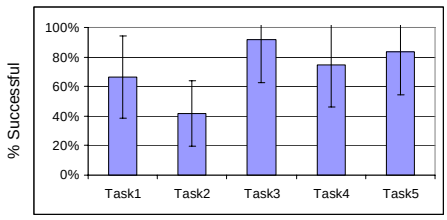
= AVERAGE(F2:F13)

= Calculated based on binomial distribution

5

Task Success

- Most common way to present is based on individual task
- Also look at binary success based on user or type of user
  - Frequency of use (infrequent vs. frequent users)
  - Previous experience with product
  - Domain expertise (low- vs. high-domain knowledge)
  - Age group



Average success rate for Task 1 is 67%

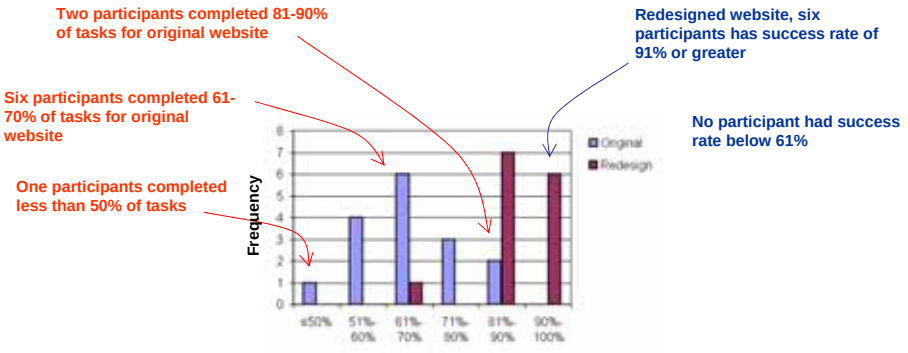
There is a 95% chance that the true mean is between 39-95%

Important to provide confidence intervals!

6

Task Success

- Looking at success by participant
  - Able to calculate percentage of tasks successfully completed
  - Data no longer binary, it becomes continuous
- Frequency distribution
  - Convenient way to visually represent variability in binary success data



7

Task Success

- Levels of Success
  - Useful when shades of gray associated with task success
  - Participant receives some value for partially completing a task
  - Valuable to know why some participants failed to complete a task, or with which tasks they needed help

8

Task Success

- Similar to measuring binary success, except you need to define the various levels
  - Extend or degree a participant completed a task
  - Received any assistance or got only part of the answer
  - Experience in completing a task (struggled vs. no difficulty)
  - Accomplishing task in different ways (optimal vs. non-optimal)

Six level of completion method

Complete success (score = 1.0)
- ☐ With assistance
- ☐ Without assistance

Partial success (score = 0.5)
- ☐ With assistance
- ☐ Without assistance

Failure (score = 0)
- ☐ Participant thought it was complete, but it wasn't
- ☐ Participant gave up

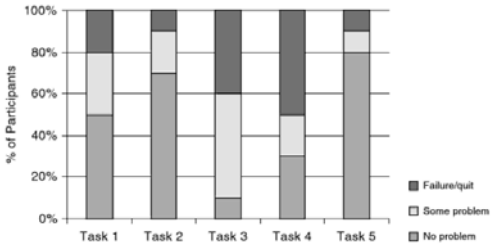
4-point scoring method

- ☐ 1 = No problem. Participant successfully completed the task without any difficulty or inefficiency.
- ☐ 2 = Minor problem. Participant successfully completed the task but took a slight detour
- ☐ 3 = Major problem. Participant successfully completed the task but had major problems. S/he struggled and took a major detour in their eventual successful completion of the task
- ☐ 4 = Failure/gave up. Participant provided the wrong answer or gave up before completing the task, or the moderator moved on to the next task before successful completion

9

Task Success

- Remember the data is ordinal
  - Don't provide average
  - Present as frequencies of each level of completion
  - Create stacked bar chart to show percentage of users that fall into each category
  - Present a usability score by assigning success range from 0 to 1, be sure to change y-axis to "average success score" instead of "% success"



10

Task Success

- Issues in Measuring Success

- How do you define task success?
  - Clearly define what criteria is for successfully completing task
  - What if user finds right answer but reports it in the wrong format?
  - What if user finds right answer but restates it incorrectly?
  - Make note of unexpected situations and try to reach consensus afterwards
- How or when to end task if participant is not successful?
  - Tell participant at beginning of session that they should continue to work on task until they complete or reach the point they would give us and seek assistance
  - Apply "three strikes and you're out rule"
    - Three attempts to complete task before you stop them
    - Difficulty lies in defining "attempt" – three strategies, detours, or wrong answers
  - "Call" task after predefined time has passed

11

Time-on-Task

- Referred to as task completion time or task time
  - Time it takes a participant to perform a task
- Good way to measure usability of product
  - In most situations the faster the participant can complete the task, the better
  - Uncommon for participant to complain that task took less time to complete than expected
- Importance of Measuring Time-on-Task
  - Task repeated by user
  - Example
    - Customer service for airline
    - Faster phone reservation completed, more calls can be handled
    - More money save

Exceptions?

Games – game experience more important than completion time

Online Training Course – better for participants to spend more time completing tasks

12

Time-on-Task

- Time elapsed between start and end of task
- How to Collect and Measure Time-on-Task
  - Use stop watch or other digital device that can measure at the minute and second levels
  - Use clock to record start and end times (helpful to have two people record times)
  - Videotape session and use the time-stamp feature
  - Automated tools
    - Ergo Browser, Data Logger, Bailey's Usability Testing Environment (UTE)
    - Can calculate average task completion times
    - Less error prone
    - Less obtrusive (user won't see you time them)
- Turning on and off the clock
  - Need rules about how to measure time
  - Have participant read aloud task, turn clock on as soon as they finish reading
  - When participants hit "answer" button, turn clock off (automated approach)
  - When participants verbally report answer / write answer down, turn clock of (manual approach)

At start of response or end of response?

As soon as interaction with product is complete

13

Time-on-Task

- Analyzing & Presenting Time-on-Task Data
  - Arrange data in table
  - Show summary data
    - Average
    - Median
    - Geometric Mean
    - Confidence Intervals (assuming 95%)

Participant	Task 1	Task 2	Task 3	Task 4	Task 5
P1	259	112	135	58	8
P2	253	64	278	160	22
P3	42	51	60	57	26
P4	38	108	115	146	26
P5	33	142	66	47	38
P6	33	54	261	26	42
P7	36	152	53	22	44
P8	112	65	171	133	46
P9	29	92	147	56	56
P10	158	113	136	83	64
P11	24	69	119	25	68
P12	108	50	145	15	75
P13	110	128	97	97	78
P14	37	66	105	83	80
P15	116	78	40	163	100
P16	129	152	67	168	109
P17	31	51	51	119	116
P18	33	97	44	81	127
P19	75	124	286	103	236
P20	76	62	108	185	245
Average	**86.6**	**91.5**	**124.2**	**91.35**	**80.3**
Median	**58.5**	**85**	**111.5**	**83**	**66**

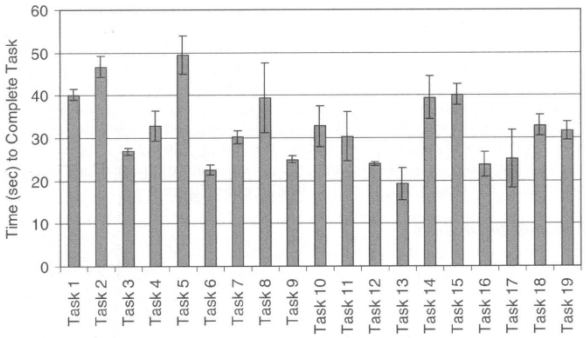
Participant	Task 1	Task 2	Task 3	Task 4	Task 5
Geometric mean	65.216	85.225	104.971	73.196	60.323
Upper bound	119.8	108.0	159.5	116.6	110.2
Lower bound	53.4	75.0	119.9	66.1	50.4
Confidence interval	33.2	16.5	19.8	25.2	29.9

Note: Data are all expressed in seconds.

14

Time-on-Task

- Averages
  - Most common way to present task-on-time is to look at average time on any task or set of task
  - Variation across individuals can impact average time
    - 95% confidence interval provided
  - See variability within same task as well as across tasks

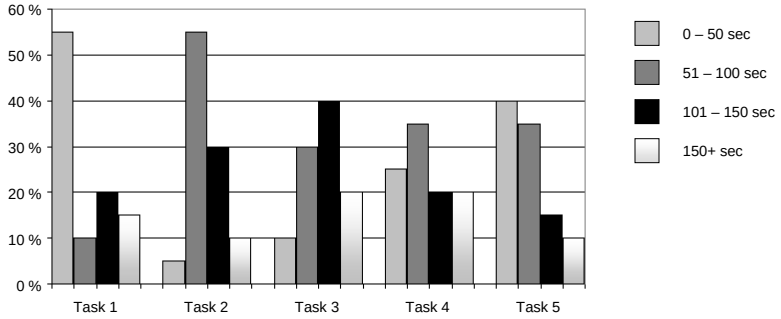


Note: Data from a different online study of prototype website (not the 5 tasks from previous slide)

15

Time-on-Task

- Ranges
  - Create ranges or discrete time intervals
  - Report frequency of participants who fall into each
  - Able to look for patterns in the type of participants who fall into certain categories
    - Do participants with really long completion times have common characteristics?

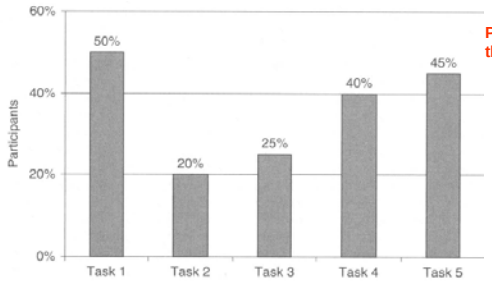


Note: Figure utilizes data from slide 14

16

Time-on-Task

- Thresholds
  - Many cases, only matters that user can complete task within an acceptable amount of time
  - Average may be unimportant, mail goal is to minimize the number of users who need excessive amount of time
- What should the threshold be?
  - Do task yourself (assuming you are an expert), double that time
  - Based on competitive data
  - Best guess



Percentage of users below the threshold (1 minute)

Note: Figure utilizes data from slide 14

Time-on-Task

- Distributions or Outliers
  - Critical to look at distribution, particularly for automated tools or when moderator not present
  - Task time of 2 hours compared to 15 to 20 seconds indicates problems
  - Acceptable to exclude outliers from analysis
    - Statistical techniques for identifying outliers
    - Book suggests anything more than three standard deviations above the mean
    - Below mean causes same problem, have expert run through task (minimum acceptable time) and anything below is thrown out

Participant	Task 1
P1	259
P2	253
P3	42
P4	38
P5	33
P6	33
P7	36
P8	112
P9	29
P10	158
P11	24
P12	108
P13	11
P14	37
P15	116
P16	129
P17	31
P18	33
P19	75
P20	76
Average	81.65
Median	40

standard dev = 72.57

outlier = 81.65 + (3 x 72.57)

outlier = 299.36

expert time = 25 sec

Time-on-Task

- Issues to Consider when Using Time Data
  - All tasks vs. only successful tasks
- All tasks?
  - Advantage of using only successful task it's a good measure of efficiency
  - Difficult to estimate time for unsuccessful tasks, some participants keep trying until you "pull the plug"
- Only successful tasks
  - Advantage of using all data is that it's more accurate measure of user experience
  - Using all data makes it independent, using only successes make it dependent on success rate
- Rule of thumb
  - Participant give up, use data
  - Moderator pulls plug, use only success times

19

Time-on-Task

- Issues to Consider when Using Time Data
  - Impact of thinking aloud protocol
- Retrospective probing techniques
  - Thinking aloud provides important insight into user experience
  - Some users provide long speech on the important of fast-loading webpage (or whatever) which impacts time-on-task
  - Ask user to "hold" comments until task complete, then have dialog
- Real impact?
  - Some argue that thinking aloud decreases time-on-task by helping participants focus on task, organize how to accomplish task, recover from error

20

Errors

- Usability professionals believe errors and usability issues are the same thing
  - Are they?
- Usability issue – underlying cause of a problem
  - User experience problem completing purchase on website
  - Issue or cause may be the confusing labeling of products
- Error – one or more possible outcome
  - User experience problem completing purchase on website
  - Error or result of the issue is the act of choosing the wrong option for the product they want to buy
  - Errors are incorrect actions that may lead to failure

21

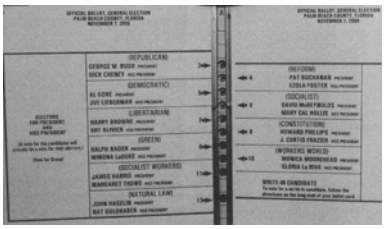
Errors

- When to Measure Errors
  - Helpful to classify errors instead of just documenting usability issues
  - Able to understand specific action or set of actions that result in task failure
- General situation where measuring errors is useful
  - Error result in significant loss in efficiency
    - Results in loss of data
    - User needs to re-enter data
    - Significantly slower task completion time
  - Error results in significant cost
    - Result in increased call volumes to customer service
    - Increased product returns
  - Error results in task failure
    - Cause patient to receive wrong medication
    - Voter accidently votes for wrong candidate
    - Web user buys wrong product

22

Errors

- What Constitutes an error?
  - No widely accepted definition
  - Some type of incorrect action on the part of the user that prevents the user from completing the task in the most efficient manner
- Types of actions errors can be based on
  - Entering incorrect data in a form field
  - Making the wrong choice in a menu or drop-down list
  - Taking an incorrect sequence of actions
  - Failing to take a key action



Infamous "butterfly ballot" used in 2000 presidential election

Record vote by punching one of the holes in center strip

Al Gore is second candidate listed on the left, to vote for him need to punch third hole

23

Errors

- Collecting and measuring errors
  - Need to know correct action or correct set of actions
  - Does a task present a single error opportunity or are there multiple error opportunities?
  - If multiple error opportunities, do you care about all of them or only one of them?
- Organizing error data
  - For single errors, use 1/0
  - For multiple errors, record number of errors for each task and user
  - Not all errors are equal, weight each type of error

Single error opportunities
1 = Error, 0 = No Error

Task 1	0
Task 2	1
Task 3	1

Multiple error opportunities
Count number of errors

Task 1	0
Task 2	4
Task 3	2

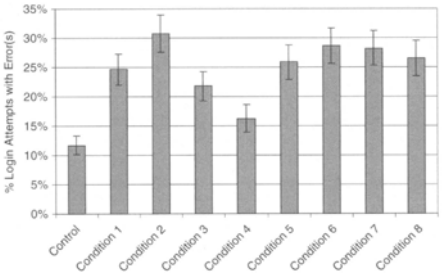
Multiple error opportunities -Severity ratings
1 = Trivial, 2 = Moderate, 3 = Serious

Task 1	0
Task 2	1
Task 3	3

24

Errors

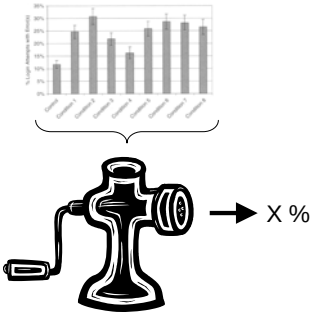
- Analyzing and presenting errors – Single Error Opportunities
  - Look at frequency of the error for each task
- Option 1
  - Plot out number of error
  - Shows number of error for each task, no need for confidence intervals
- Option 2 (shown on right)
  - Divide total number of errors by total number of participants
  - Percentage of participants who made an error for each task
  - Useful if there are different number of participants for each task



25

Errors

- Analyzing and presenting errors – Single Error Opportunities
  - Aggregate perspective
- Option 1
  - Average into a single error rate
  - Overall error rate for study = 25%
  - Useful for bottom line metric reporting
- Option 2
  - Average of all tasks that had a certain number of errors
  - 50% of all tasks had an error rate of 10% or greater
  - At least on participant made an error on 80% of the tasks
- Option 3
  - Establish maximum acceptable error rate for each task (e.g. 10%)
  - 25% of tasks exceeded an acceptable error rate



26

Errors

- Analyzing and presenting errors – Multiple Error Opportunities
- Option 1
  - Frequency of errors for each task
  - May be misleading if different number of error opportunities
  - Divide by total number of error opportunities
- Option 2
  - Calculate average number of error made by each participant
  - Indicates which tasks produce most errors
  - Suggest number of errors a typical user may face
- Option 3
  - Tasks that fall above/below a threshold
- Option 4
  - Weight error by severity
  - Add up "error score" for each participant, divide by total number of participants
  - Different than error rate – indicates which task have more frequent and/or serious errors

http://media.photobucket.com

27

Errors

- Issues to Consider When Using Error Metrics
- Don't double count errors
  - User typed in an extra character in the password field
  - If counting "extra character" as an error, don't also count as "incorrect character"
- Need to know more than just error rate
  - Why are different errors occurring?
  - Try to code each type of error
    - missing character, extra character, navigation error, selection error, interpretation error
  - Able to better understand where problems are
- Error may be the same as task failure
  - Error on a login page is also task failure
  - Could just report task failure

28

Efficiency

- How to measure?
  - Time-on-task is a measure of efficiency
  - Also look at from amount of effort required to complete task
    - Number of steps needed to perform task
    - Most product want to minimize effort by reducing the number of discrete events required
- What is effort?
  - Cognitive effort – finding right place to perform an action
    - Finding link on webpage
    - Deciding what action is necessary
    - Interpreting results of the action
  - Physical effort – physical activity required to take action
    - Moving the mouse
    - Inputting text on a keyboard
    - Turning on a switch

Automobile navigation system

Minimize both cognitive and physical effort required

29

Efficiency

- Collecting and Measuring Efficiency
  - Identify the actions to be measured
    - Mouse clicks, page views, keystrokes, button presses
  - Define the start and end of an action
    - Duration varies
    - Actions can be passive
  - Count the actions
    - Actions must happen at a pace that can be identified visually
    - Use automated system if needed, avoid having to watch hours of video
  - Action must be meaningful
    - Each action should represent an incremental increase in cognitive and/or physical effort
    - The more actions, the more effort
  - Look only at successful tasks
    - Participant may only take a few steps and quit
    - Looks like this participant was very efficient

30

Efficiency

- Analyzing and Presenting Efficiency Data
- Average
  - Look at number of actions each participant takes to complete a task
  - Calculate average for each task
  - Don't forget the confidence interval
- Lostness
  - Metric used in studying behavior on the web

N : The number of different web pages visited while performing the task

S : The total number of pages visited while performing the task, counting revisits to the same page

R : The minimum (optimum) number of pages that must be visited to accomplish the task

$$L = \text{sqrt}\,[(N/S-1)^2 + (R/N-1)^2]$$

31

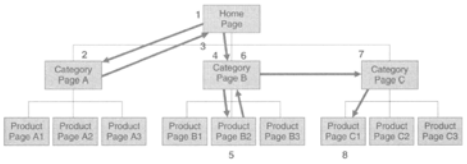
Efficiency

$L = \text{sqrt}\,[(N/S-1)^2 + (R/N-1)^2]$

N : The number of different pages visited
S : The total number of pages visited
R : The minimum number of pages

- Lostness Example - Participant's task is to find something on Product Page C1

$L = \text{sqrt}\,[(3/3-1)^2 + (3/3-1)^2]$
$= \text{sqrt}\,[0 + 0]$
$= \text{sqrt}\,[0]$
$= 0$



$L = \text{sqrt}\,[(6/8-1)^2 + (3/6-1)^2]$
$= \text{sqrt}\,[0.0625 + 0.25]$
$= \text{sqrt}\,[0.3125]$
$= 0.56$

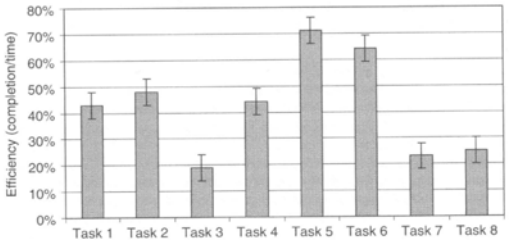
Lostness < 0.4 – participants did no exhibit characteristics of being lost
Lostness > 0.5 – participants appeared to be lost

32

Efficiency

- Combination of task Success and time
  - Common Industry Format (CIF) for Usability Test Reports specifies the “core measure of efficiency” is the ratio of the task completion rate to the mean time per task

Task	Completion Rate Percentage		Task Time (mins)		Percent Efficiency
1	65		1.5		43
2	67		1.4		48
3	40		2.1		19
4	74		1.7		44
5	85		1.2		71
6	90		1.4		64
7	49		2.1		23
8	33	/	1.3	=	25



Higher values of efficiency are better

Task 5 & 6 appear more efficient than others

33

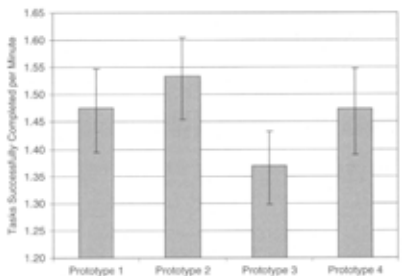
Efficiency

- Combination of task Success and time
  - Variation is to count number of tasks successfully completed by each participant
  - Divide by the total time spend on all tasks (successful and unsuccessful)

Participant completed 10 tasks successfully

Total time 10 minutes

1 task successful per minute



Between subjects study comparing four different navigation prototypes for a website

Participant asked to use on of 4 prototypes to perform same 20 tasks

Each prototypes tested by more than 200 participants

Counted number of tasks successfully completed by each participant and divide by the total time each participant spent

34

Learnability

- Most products require some amount of learning
- Learning happened over time as experience increases
  - Based on the amount of time spent using a product and variety of tasks performed
- Learnability
  - Extent to which something can be learned
  - Measured by looking at how much time and effort required to become proficient with something

35

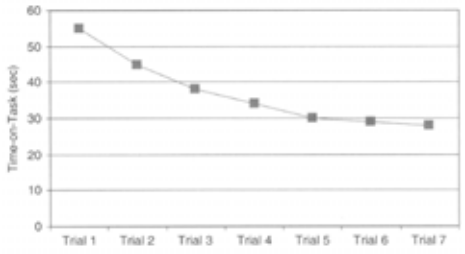
Learnability

- Collecting and Measuring Learnability Data
  - Data collected at multiple times
  - Each instance of collecting data is a trial
  - Trial can be every five minutes, every day, or once a month (based on expected frequency of use)
- Any performance metric over time can be used to measure Learnability
  - Time-on-task
  - Errors
  - Number of steps
  - Task success per minute
- Types of trials
  - Within the same session
    - Easy to administer, does not take into account memory loss
  - Within the same session but with breaks between tasks
    - Break can be a distracter task, or anything that promotes forgetting
    - Easy to administer, session become relatively long
  - Between sessions
    - Least practical, but most realistic

36

Learnability

- Analyzing and Presenting Learnability Data
  - Examine a specific performance metric by trial for each task (or set of tasks)
  - Performance will change as a function of experience



Slope of the line sometimes called the learning curve

Determine whether statistically significant difference between learning curves, analyze variance

Notice the point of **asymptote** – where the line starts to flatten out

How long to reach maximum performance?

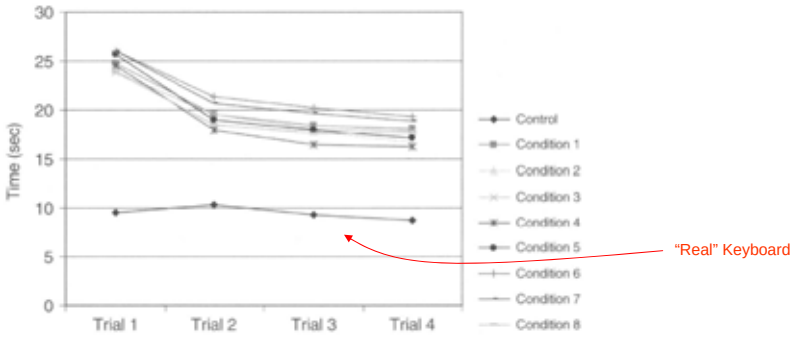
Difference between highest and lowest values on y-axis

How much learning needed to reach maximum performance?

37

Learnability

- Analyzing and Presenting Learnability Data
  - Compare Learnability across different conditions



How speed (efficiency) of entering password changed over time using different types of on-screen keyboards

38

Learnability

- Issues to Consider When Measuring Learnability
- What should be considered a trial?
  - Learning can be continuous
  - Learning is more about developing and modifying different strategies to complete a set of tasks
  - Trials don't make sense, take measurements every 5, 15, or 60 minutes
- How many trials to include?
  - At least two, most cases should really be at least 3 or 4
  - Err on the side of more trials than you think to reach stable performance

39

Summary

- Performance metrics evaluate usability of any product, five general types
  - Task success
    - Interested where participant able to complete task
    - Binary or based on degree of completion, the experience in finding an answer, or the quality of answer
  - Time-on-task
    - Concerned about how quickly users can perform tasks with product
    - Look at time it takes all users to complete task, portion of the users, or those who can complete within time limit
  - Errors
    - Useful measure based on number of mistakes made while attempting to complete a task
    - Single or multiple error possibilities, differing levels of importance
  - Efficiency
    - Amount of effort (cognitive or physical) required to complete a task
    - Number of steps or actions to complete task or ratio of task success to average time per task
  - Learnability
    - Looking at how efficiency metrics change over time
    - How and when participants reach proficiency in using a product

40