

Tutorial: Reporting Usability Data That Is Scientifically Sound

Bill Killam

Since testing is nearly always conducted with a small sample, there are restrictions on how usability test results can be reported and still be considered valid. This tutorial will demonstrate how best to report effectiveness, efficiency, and satisfaction data with or without statistical significance. Even though this is listed as an advanced tutorial for experienced practitioners, experience in calculating statistics is not a requirement.

About the Speaker:

Bill Killam is the President and Principle Human Factors Engineer at User-Centered Design, Inc. Mr. Killam has degrees in both engineering and psychology and is board certified in Human Factors Engineering by the Board of Certification in Professional Ergonomics. He has been providing Human Factors Engineering, user-centered design, and usability services for over 25 years. He and has provided web site design and testing service to the numerous US Government agencies including the State Department, the US Geological Survey, Centers for Disease Control and Prevention, Food and Drug Administration, NIST, and the National Cancer Institute as well as numerous commercial and non-profit organizations. Mr. Killam teaches as an adjunct professor at both the University of Maryland and George Mason University and has taught the workshop on usability at the University of Maryland's HCIL Open House symposium for the past 7 years. He is an active member of the human factors engineering and usability testing community at both the national and local level and has been the Vice President and President of the Potomac Chapter of the Human Factors and Ergonomics Society, president of the DC Chapter of the Usability Professionals Association, and is on the board of DC Chapter of the Association of Computing Machinery's Special Interest Group on Human Computer Interaction (ACM SIGCHI).

Reporting Usability Test Data That Is Statistically Sound

October 12, 2007

Bill Killam, MA CHFP

bkillam@user-centereddesign.com

(703) 729-0998

User-Centered Design • www.user-centereddesign.com

People don't understand statistics

- “There’s a 20% change of snow on the ground on Christmas day. And since we hadn’t had snow on the ground for 4 years, we should have it this year.”
- 30% change of rain in the DC area
- There’s as much of a chance of the lottery producing 0-0-0-0-0 as any other number
- There’s a greater chance of having a tree falling on you in a thunderstorm than being struck by lightning. But there’s a greater chance of being struck by lightning than winning the Virginia lottery

People don't understand statistics

- 4 out of 5 dentist surveyed say...
- "40% of the people who try to find data on the [whoever's] site can't find what they are looking for"
- The average number of clicks to accomplish [a task in a test] was 5
- The average time to complete a task was 20 seconds

The problem...

- Discount usability is generally performed with just a few subjects so if recruiting is not tightly controlled we have to be careful what we say and imply we can prove it
- If not, it can lead to misunderstanding
 - Contract Requirement: The selected vendor will be required to demonstrate that one design (or one version of a design) is better than another design (or version) using 8 participants per study.

Descriptive Statistics

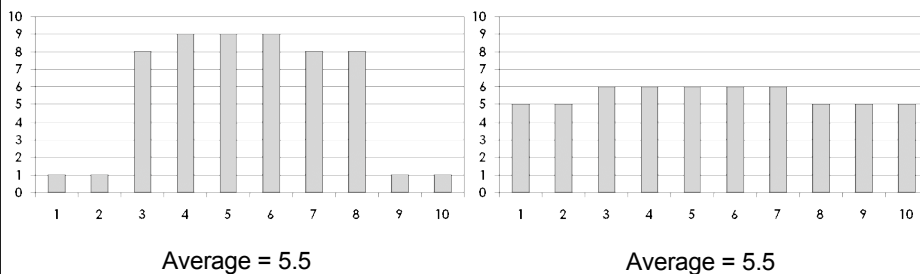
- Descriptive statistics summarize data without implying or inferring anything beyond the sample tested
 - This type of statistic can always be reported
 - This includes the mean, median, mode, variance, standard deviation, and covariance
 - But should be reported in an appropriate way

5

User-Centered Design • www.user-centereddesign.com

Descriptive Statistics

- Providing the average (e.g., 5.5) is not sufficient...



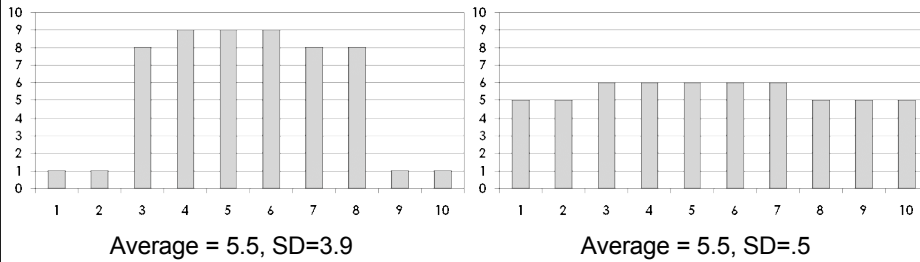
6

User-Centered Design • www.user-centereddesign.com

Descriptive Statistics

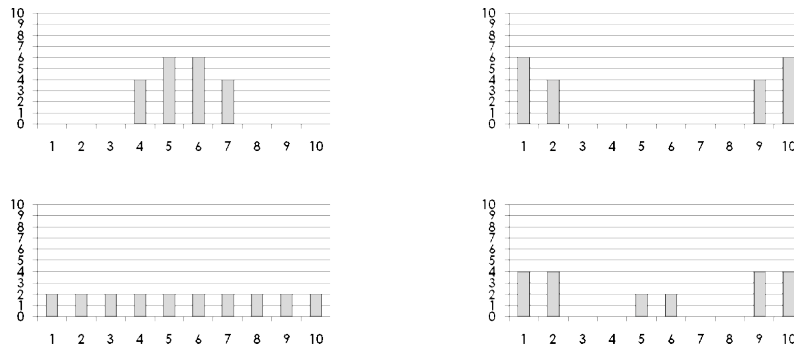
➤ Adding the standard deviation can be “helpful”

AVG: 5.5, 3.9 versus AVG 5.5, SD=.5....



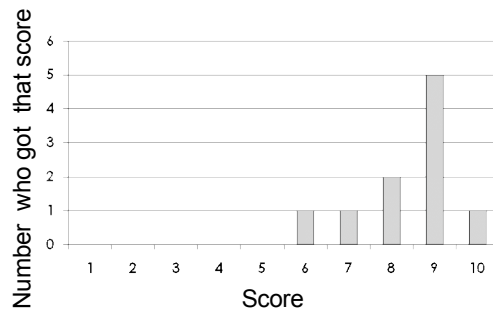
Descriptive Statistics

➤ But the data often shows other patterns such as bimodal distributions. In these cases, the average and standard deviation are not adequate...



Histograms

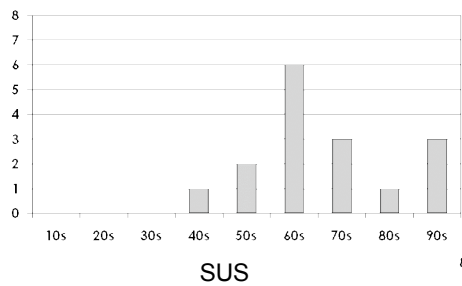
- You could report the mode, which has value by itself (e.g., 6 of 9 people did x), but a histograms is probably the best way to present the data since it shows the most data of interest...



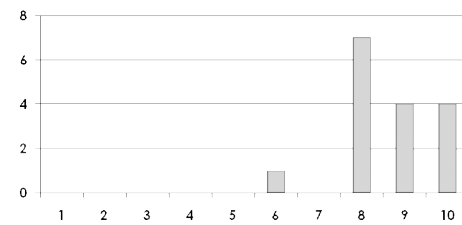
9

User-Centered Design • www.user-centereddesign.com

Histograms (continued)_



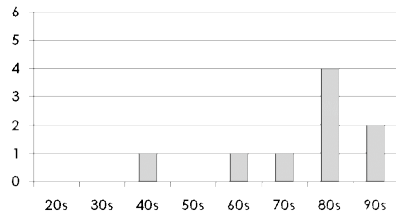
SUS



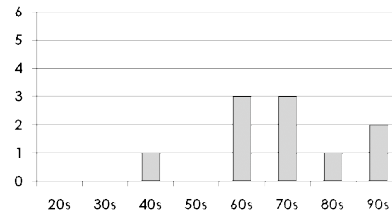
Cooper Harper

User-Centered Design • www.user-centereddesign.com

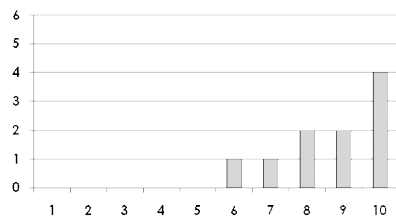
Histograms (concluded)_



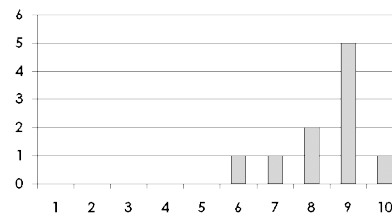
Memphis – SUS Data



DC – SUS Data



Memphis – MCH Data

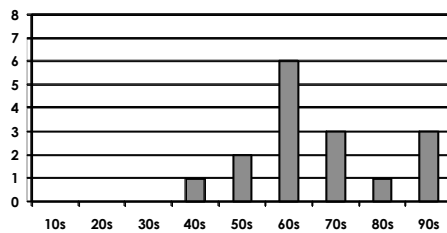


DC – MCH Data

11

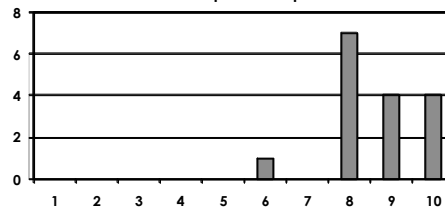
User-Centered Design • www.user-centereddesign.com

Reporting the Results (continued)



SUS Data

Cooper Harper



MCH Data

User-Centered Design • www.user-centereddesign.com

Predictive Statistics

- Predictive (or inferential statistics) is a means drawing conclusions about a population based on data from a sample population
- This type of statistic can only be reported when sampling is conducted that is truly representative of the underlying population and the procedures must be capable of drawing correct conclusions about the population
- To use predictive statistics, the test must be both valid and reliable

Validity

- Validity is the degree to which the results of a research study provide trustworthy information about the truth or falsity of the hypothesis (from Cherulnik, P.D. (2001). *Methods for Behavioural Research: A Systematic Approach*)

Validity

- **Internal validity** refers to the situation where “experimental treatments make a difference in this specific experimental instance” (from Cambell, D.T. & Stanley, J.C. (1963) *Experimental and Quasi-experimental Designs for Research*
 - Failure to counter balance
 - Learning curve effects (ease-of-learning versus ease-of-use)_
 - Disruptive protocols
- **External validity** asks the question of “generalizability”
 - Sample size issues
 - Representative issues
 - Nonrepresentative research context (e.g., the Hawthorne effect)_

Validity

- **Construct validity issues**
 - **Content validity** and **face validity** are closely related. Content validity is an assessment that the test is valid as assessed by experts. Face validity is an assessment that the test is valid as assessed by the casual observer (a non expert)_
 - **Inadequate operationalization** refers to a failure to define what will be measures in a way that is necessary and sufficient
 - **Measurement artifacts** refers to measurement biases such as experimenter expectancy or strategic responding during think aloud protocols
 - **Concurrent validity** refers to the relationship between this test and other tests that purport to measure the same data

Reliability

- Reliability is the ability of a test to show the same results if conducted multiple times
 - Test-retest reliability
 - Repeatability
 - Reproducibility

Use of Confidence Intervals

- “You just finished a usability test. You had 5 participants attempt a task in a new version of your software. All 5 out of 5 participants completed the task. You rush excitedly to tell your manager the results so you can communicate this success to the development team. Your manager asks, ‘OK, this is great with 5 users, but what are the chances that 50 or 1000 will have a 100% completion rate?’ ” - *Jeff Sauro*
- The confidence level tells the informed reader the likelihood that another sample will provide the same results. In other words, if you ran the test again, what value are you likely to get next time?

Use of Confidence Intervals (concluded)_

- The confidence interval is determined (in large part) by the sample size. The more people you test, the more variables you have included, so the better your confidence (i.e., the smaller the confidence interval)_
- Since time and cost are directly proportional to the sample size, there is a trade off between the margin or error/confidence interval and the cost of the evaluation

Effectiveness Data

- Effectiveness data can be operationalized in a number of ways but is generally operationalized as success or failure to complete a task
- Completion rate as a pass/fail criteria can be measured objectively if the criteria is pre-determined and is not subjective
- Best estimates, error rate, and the confidence interval can be calculated easily for pass/fail measure of completion rate using a Binomial calculation

Completion Rate (concluded)_

- Example:
 - If 4 out of 8 complete a task, you can say, with 95% certainty, that somewhere between 21% to 78% of all users will complete the task
 - If 8 out of 8 complete a task, you can only say, with 95% certainty, that somewhere between 71% to 100% of all users will complete the task
 - If 7 out of 8 complete the task, its 51%-99%
 - If 6 out of 8 complete the task, its 40%-93%
- If you want to state that 90% of the users will complete a specific task (+/-5%), and all conditions of validity and reliability have been met...
 - You need to test 156 people and 142 of them complete the task
 - (For a confidence interval of 5% (+/- 2.5%, you'd need 450 people)_

Efficiency Data – Time on Task

- Efficiency data can be operationalized in number of ways – time on task being the most common
- Time on task can be measured objectively
- External time is important to management, but is not necessarily important users and time on task does not correlate with effectiveness (except in extreme cases)_
- Though the data is not necessarily normally distributed, even with fairly large samples, a t-test can be used to complete time provided it is valid
 - It is not valid in think aloud or exploratory protocols
 - It can be invalidated by sampling issues

Satisfaction Data

- Satisfaction data can be operationalized in a number of ways, but is always opinion data
 - Standardized survey instrument (e.g. SUS)_
 - Simple Likert scale assessments
- Questionnaires suffer from numerous issues that threaten their validity
 - Halo effect, leniency effect, strictness effect, bias
- Satisfaction data does not correlate with performance except in extreme situations

Satisfaction Data (continued)_

- Take a poll on participants comparing the new product against the old version of the product. People might be asked to comment on the statement 'The new design is an improvement over the old design.' and given a choice of answers from "Definitely, it's the tops." to "No definitely not, it's awful." The data would be a collection of opinions. Assume the following scale and results...

Person Number	<i>Definitely an improvement, it's the tops.</i>	<i>It's a good improvement</i>	<i>it's OK</i>	<i>I have no opinion</i>	<i>Not much of an improvement.</i>	<i>I don't think its an improvement</i>	<i>No, definitely not an improvement, it's awful</i>
1					X		
2		X					
3	X						
4				X			
5					X		
6				X			
7	X						
8						X	
9			X				
10		X					
11			X				
12	X						

Satisfaction Data (continued)_

- Though it appears to be one, this is not an interval scale since you cannot state that an opinion rated as 3 is three times weaker than the opinion rated as 1. There is, however, an order to the different opinions (its an ordinal scale).
- Descriptive statistics (average, standard deviation) cannot be calculated, nor can parametric statistics be used to analyzed the results since the data is not normally distributed
- You can use a non parametric analysis such as a rank sum analysis

Rank Sum Analysis

- The hypothesis you would want to test would be: “The participants consider the new product an improvement.”
- A quick ‘eyeball’ test shows that none of those questioned thought it was awful and only one person thought it not very good, so a first impression is that people generally approve. If you start by assuming that in the population there is no opinion one way or the other, and that people’s responses are symmetrically distributed about ‘no opinion’, you can test the hypothesis that people think the shopping center is an asset, with the null hypothesis that people have no opinion about it, their response being the median value 4.

Rank Sum Analysis (continued)_

- You need to be careful to choose the appropriate test statistic for the problem you are tackling
 - For a **one** tailed test, where the alternative hypothesis is that the median is **greater** than a given value, the test statistic is W^- . For a **one** tailed test, where the alternative hypothesis is that the median is **less** than a given value, the test statistic is W^+ .
 - For a **two** tailed test the test statistic is the smaller of W^+ and W^-
- As people who think it an improvement will give a rating of less than 4, the null and alternative hypotheses can be stated as follows.
 - H_0 : the median response is 4
 - H_1 : the median response is less than 4
 - 1 tail test, Significance level 5%

User-Centered Design • www.user-centereddesign.com

Rank Sum Analysis (continued)_

- List the value
- Find the difference between each value and the median.
- Ignore the zeros and rank the absolute values of the remaining scores.
- Ignore the signs, start with the smallest difference and give it rank 1. Where two or more differences have the same value find their mean rank, and use this.
- Now check that $W^+ + W^-$ are the same as $\frac{n(n+1)}{2}$, where n is the number in the sample (having ignored the zeros). In this case $n = 10$.
 - $\frac{n(n+1)}{2} = \frac{10 \times 11}{2} = 55$
 - $W^+ + W^- = 9.5 + 45.5 = 55$

rating	rating - median (4)	absolute value	ranking	+	-
5	5 - 4 = 1	1	2	2	
2	2 - 4 = -2	2	5.5		5.5
1	1 - 4 = -3	3	9		9
4	4 - 4 = 0	0	Ignore		
5	5 - 4 = 1	1	2	2	
2	2 - 4 = -2	2	5.5		5.5
4	4 - 4 = 0	0	Ignore		
1	1 - 4 = -3	3	9		9
6	6 - 4 = 2	2	5.5	5.5	
3	3 - 4 = -1	1	2		2
2	2 - 4 = -2	2	5.5		5.5
1	1 - 4 = -3	3	9		9
Total				9.6	45.5

User-Centered Design • www.user-centereddesign.com

Rank Sum Analysis (concluded)

- Compare the test statistic with the critical value in the tables. If the null hypothesis were true, and the median is 4, you would expect W_+ and W_- to have roughly the same value. There are two possible test statistics here, $W_+ = 9.5$ and $W_- = 45.5$, and you have to decide which one to use. We are interested in W_- , the sum of the ranks of ratings greater than 4. W_+ is much less than W_- which suggests that more people felt the shopping center was an asset. It could also suggest that those who expressed a negative view expressed a very strong one, with lots of high numbers in the ratings.
- Now you need to compare the value of W_+ , the test statistic, with the critical value from the table. Given that W_+ is small the key question becomes "Is W_+ significantly smaller than would happen by chance?" The table helps you decide this by supplying the critical value. For a sample of 10, at the 5% significance level for a 1 tailed test, the value is 10. As W_+ is 9.5, which is less than this, the evidence suggests that we can reject the null hypothesis.
- Your conclusion is that the evidence shows, at the 5% significance level, that the public thinks the design is better than the old design

1-tail	5%	2_%	1%	_%
2-tail	10%	5%	2%	1%
n				
2	-	-	-	-
3	-	-	-	-
4	-	-	-	-
5	0	-	-	-
6	2	0	-	-
7	3	2	0	-
8	5	3	1	0
9	8	5	3	1
10	10	8	5	3
11	13	10	7	5
12	17	13	9	7
13	21	17	12	9
14	25	21	15	12
15	30	25	19	15

User-Centered Design • www.user-centereddesign.com

Summary

- With discount usability, or any exploratory approach, any predictive statistic calculated will not be valid (or will not be worth the effort)_
- Descriptive statistics are valid with small numbers, but Histograms are the best way to show data from small samples

Summary

- The “Real” data from small number testing comes from the “Principle of Inter-ocular Drama” and the observer’s ability to explain it
- Effectiveness, efficiency, and satisfaction are not correlated with each other. Only effectiveness is correlated with the root definition of usability so efficiency and satisfaction should only be used to add supplementary information after acceptable levels of effectiveness has been proven or presumed to be acceptable