Using a Single Usability Metric (SUM) to Compare the Usability of Competing Products

Jeff Sauro

Oracle, Inc.
Denver, CO, USA
Jeff.Sauro@oracle.com

Erika Kindlund

Intuit, Inc.
Mountain View, CA, USA
Erika_Kindlund@intuit.com

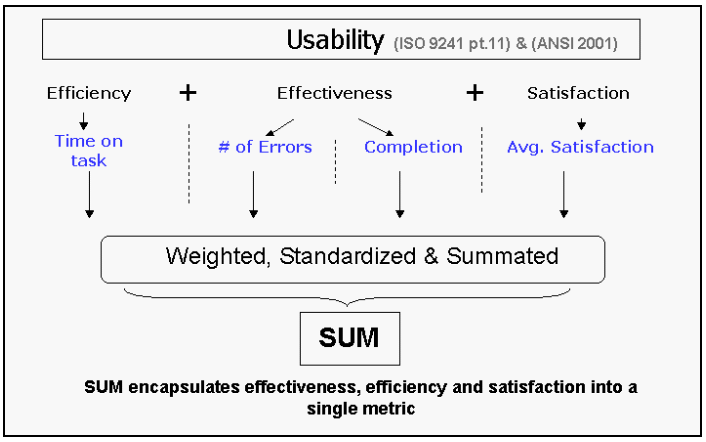
Abstract

This paper provides additional data and examples to substantiate a method to simplify all the ANSI and ISO aspects of usability into a single, standardized and summated usability metric (SUM). Using data from three additional data sets and in the context of a competitive usability test, differences between SUM scores are compared to conventional raw usability metrics. We demonstrate how SUM can be an effective way to summarize the usability of a product or tasks to business leaders and how it provides a better way for making decisions about usability.

1 INTRODUCTION

SUM is a single, summated and standardized metric originally introduced in [10] that encapsulates the majority of information in four common usability metrics. It is based on the ANSI [1] and ISO 9241 pt.11 [4] defined dimensions of usability—effectiveness, efficiency & satisfaction. The four metrics that are used to derive the SUM score of a system in a summative evaluation are: task completion rates, average number of errors, average time on task, and post-task satisfaction. This summative model of usability is illustrated in Figure 1. These four component metrics are combined using standardization methods adapted from Six Sigma [2].

Figure 1: SUM Model



Earlier research across four summative evaluations from 4 products and 1860 task observations found that the four usability metrics had a moderate and significant correlation and an average of the standardized values would provide the maximum amount of information in one score [10]. An additional 1085 task observations with three additional products have since been examined and similar results were found. Table 1 below displays the correlations between the four metrics from the combined data set of 2940 observations. The seven data sets are made up of 238 unique users, 88 tasks and 6 different products.

Table 1: Correlation Matrix of 2940 observations from 7 summative usability evaluations on 6 unique products, from 238 users and 88 tasks.

	Time	Errors	Satisfaction
Errors	.553		
Satisfaction	-.324	-.334	
Completion	-.297	-.411	.381

All Correlations significant at the $p < .001$ level

The addition of the these three data sets provides more evidence that these common usability metrics correlate and the average of their standardized values will provide the majority of information in a single number. This single usability metric is a better estimate of "usability" than any one of the individual metrics and provides an easier way to describe the usability of a task or system.

1.1 SUM Scores & Competitive Analysis

In addition to analyzing additional data sets from new products, this paper also examines the usage of SUM as a method to effectively make business decisions about the usability of a product or task in the context of a competitive analysis. In theory, SUM scores should decrease the complexity of identifying differences in usability of competing products by summarizing the information of the four component metrics into a single measure. Using this single measure makes comparisons between products easier than reviewing the individual component metrics. Business leaders are more apt to consider the impact of usability when it is represented as a single metric along-side other competitive analysis data such as market penetration, net promotion scores [9] and cost.

2 Method

To illustrate the advantage of a single usability metric over multiple component metrics in making business decisions, we conducted a competitive summative usability test between three commercially available desktop software applications (Product X, Product Y and Product Z). Such a scenario may exist when an organization is considering purchasing competing products or when usability engineers need a benchmark of usability for a product redesign. The goal of the test was to determine which product is more "usable." While such broad statements of usability are often made about products, testing the veracity of such claims is not as easy. We chose 10 "core" tasks that touched upon the major pieces of functionality in the products to assess the overall product usability.

A between-subjects test was designed to collect the four component metrics. 49 participants were asked to complete 10 tasks on Product X, 32 different participants attempted to complete the same 10 tasks on Product Y and 21 participants attempted the same tasks on Product Z. Participants were screened and selected to provide a spectrum of product usage with varying degrees of product expertise, domain knowledge, age, gender and general computer experience. All tests were conducted in the US. A post-task question also verified product experience and no significant differences existed in overall product usage.

Tasks were counterbalanced to minimize sequence effects. For each participant, the four component metrics were collected for each task: task success/failure, number of errors committed, time on task, and average task satisfaction. SUM scores were derived from the raw data (see section 2.1 below).

After each task, all participants were asked to complete a post task questionnaire containing 5-point semantic distance scales with the end points labeled (e.g. 1:Very Difficult to 5:Very Easy). For the analysis we created a composite satisfaction score by averaging the responses from questions of overall ease, satisfaction and perceived task time (See Table 2). The three questions had high internal-reliability (coefficient alpha > .85). The average of the responses (instead of the response from only one question) provided a less error-prone score and one more descriptive of the users' perceived sense of usability, see [6 esp p.15] and [8].

Table 2: Post-task Questionnaire

How would you describe how difficult or easy it was to complete this task? Very Difficult Very Easy 1 2 3 4 5
How satisfied are you with using this application to complete this task? Very Unsatisfied Very Satisfied 1 2 3 4 5
How would you rate the amount of time it took to complete this task? Too Much Time Very Little Time 1 2 3 4 5

2.1 Calculating SUM Scores

After the raw metrics were collected for each participant, SUM scores were calculated for each task using the procedures outlined in [12]. For each task, the resulting raw data was standardized and summated into a SUM score for both products by task.

2.1.1 Standardized Task Completion

Task completion stayed as the proportion of tasks completed to total tasks attempted.

2.1.2 Standardize Task Time

The standardized task time value was calculated by subtracting the specification limit from the mean task time and dividing by the standard deviation. The percentage of area under the normal curve corresponding to the standardized (z-value) is the percent equivalent used for the standardized value.[1] The task time data was slightly positively skewed (as is often the case with time data) so the data was transformed using the natural logarithm and standardized using the logged values. The results are displayed in Table 3. Specification limits were derived using the method as described in [11].

Table 3: Mean task times, *standard deviation*, specification limits, z-value and standardized percent equivalent by task and product.

	Mean *SD*			**Spec**	**Z Value**			**% Equivalent**		
	X	Y	Z	**Limit**	X	Y	Z	X	Y	Z
Task										
1	144 *85*	94 *26*	97 *33*	131	0.02	1.35	1.07	51	91	86
2	56 *56*	64 *47*	44 *34*	81	0.98	0.74	1.49	84	77	93
3	207 *91*	293 *116*	183 *88*	253	0.68	-0.11	0.90	75	46	82
4	142 *105*	136 *73*	75 *68*	171	0.60	0.66	1.66	73	75	95
5	135 *97*	140 *77*	92 *59*	142	0.41	0.28	1.06	66	61	85
6	177 *119*	194 *114*	144 *54*	197	0.51	0.29	1.03	70	61	85
7	274 *87*	439 *133*	281 *74*	458	1.78	0.28	2.01	96	61	98
8	255 *102*	292 *113*	309 *119*	334	0.88	0.56	0.37	81	71	64
9	189 *105*	182 *90*	134 *89*	206	0.42	0.52	1.11	66	70	85
10	147 *85*	157 *85*	210 *89*	103	-0.30	-0.65	-1.33	38	26	9

2.1.3 Standardized Satisfaction

The Z-Value for satisfaction data was calculated by subtracting the specification limit (4 for all tasks) from the raw composite satisfaction score and dividing by the standard deviation. The percentage of area under the normal curve at the z-value is the percent equivalent column in Table 4. For identification of satisfaction specification limits see [7] and [12].

[1] If the reader does not have access to a table of normal values a Microsoft Excel formula can also be used. To convert a standardized z-value to the percentage of area under the normal curve, type =NORMSDIST(z-value) in any cell. To convert a percentage into a z-value (an inverse look-up) type, =NORMSINV(percentage) into any cell.

Table 4: Mean composite satisfaction scores, *standard deviation*, specification limits, z-value and standardized percentage by task and product.

	Mean *SD*			Spec	Z Value			% Equivalent		
	X	Y	Z	Limit	X	Y	Z	X	Y	Z
Task										
1	4.1 *.6*	4.0 *.5*	4.2 *.6*	4	0.12	0.08	0.36	55	53	64
2	4.3 *.7*	4.1 *.9*	4.2 *.5*	4	0.36	0.10	0.34	64	54	63
3	3.7 *1*	2.6 *1*	4.0 *.9*	4	-0.33	-1.34	-0.05	37	9	48
4	3.5 *1*	3.2 *1*	4.1 *.5*	4	-0.47	-0.77	0.15	32	22	56
5	3.8 *1*	3.5 *.8*	3.9 *.7*	4	-0.22	-0.57	-0.13	41	28	45
6	3.9 *.9*	3.7 *.9*	4.3 *.6*	4	-0.09	-0.37	0.56	46	35	71
7	3.9 *.7*	2.7 *1*	4.0 *.7*	4	-0.15	-1.29	0.05	44	10	52
8	3.9 *.7*	3.7 *.8*	3.8 *.6*	4	-0.11	-0.38	-0.27	45	35	39
9	3.3 *.9*	3.5 *.7*	3.8 *1*	4	-0.75	-0.70	-0.22	23	24	42
10	3.4 *.9*	3.3 *.9*	3.0 *1*	4	-0.63	-0.73	-1.00	26	23	16

2.1.4 Errors

The error z-value was derived by first multiplying the number of users attempting the tasks with the opportunities for an error. The total errors committed were divided by the total error opportunities to arrive at the value in the percent equivalent column in Table 5. See [12] for more discussion on identifying error opportunities.

Table 5: Total Errors, opportunities, z-value and standardized percent equivalent by task and product.

	Total Errors			# of Users			Error	% Equivalent		
	X	Y	Z	X	Y	Z	Opportunities	X	Y	Z
Task										
1	46	30	16	49	32	21	6	84	93	87
2	22	13	3	49	32	21	4	89	96	96
3	39	5	13	49	32	21	9	91	75	93
4	39	73	5	49	31	21	5	84	85	95
5	60	23	12	49	32	21	4	69	84	97
6	61	21	9	49	31	21	19	93	93	98
7	83	41	32	49	32	21	17	90	83	91
8	50	90	22	49	32	21	17	94	91	94
9	71	48	18	49	32	21	9	84	78	90
10	59	62	27	49	32	21	4	70	78	68

3 Results

3.1 Overall Product Comparison

To look for differences in usability between the products, the raw data was examined first from a product level by combining all tasks. Significance tests were performed on the raw data for the four component metrics and SUM scores in the combined data. The means and standard deviations are displayed in Table 6 below. For significant tests, a one-way ANOVA was used for task time and satisfaction scores. A Chi-Square test was used for completion rates and a non-parametric Kuskal-Wallis test was performed on the error data. A graphical representation of the means and 95% confidence intervals is also displayed in Figure 2.

Table 6: Mean and (Standard Deviation) and Ranking* for each Product by Metric and SUM.

	SUM		Completion		Time		Satisfaction		Errors	
Product X	68.8% (.08)	2	77.9% (.42)	2	172 (111)	-	3.78 (0.94)	-	1.08 (1.21)	-
Product Y	62.0% (.15)	3	68.8% (.46)	3	199 (141)	3	3.42 (1.00)	3	1.27 (1.33)	-
Product Z	76.4% (.15)	1	87.8% (.33)	1	157 (111)	-	3.93 (0.78)	-	.745 (0.86)	1

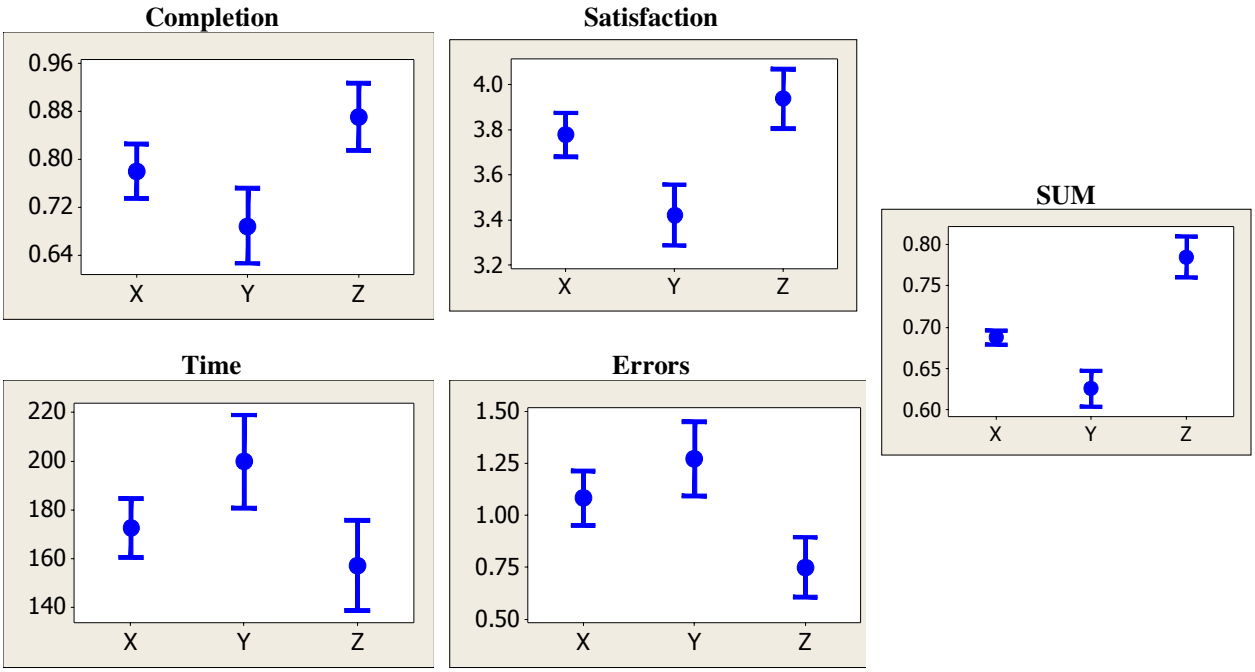
*The column to the right of the values indicate the ordering of the product's performance for the metric (1 is the best performance and 3 the worst). The "–"value indicates no significant difference at the p <.016 level.

For the combined data, all four component metrics and SUM scores show that at least one-product was significantly different than the other two: SUM scores, [$F_{.05}(2,1017) = 101.92$, $p < .001$)]; Completion [$X^2_{(2)}$ 24.16 < .001]; Time [$F_{.05}(2,1015) = 8.72$, $p < .001$)]; Satisfaction [$F_{.05}(2,1015) = 23.60$, $p < .001$)] and Errors[$H_{.05} =$ 18.19, $p < .001$].

In a three-way comparison, we do not know if all three products are significantly different from each other or their rank order of the performance. We only know that there is a significant difference with one of the products. Identifying the one product that was significantly different involves looking at the overlap of the confidence intervals. The product that has non-overlapping intervals with the other two products is the one determined to be significantly different (see Figure 2 below).

In order to determine the ranks of the three products by metric, we performed a three-way comparison between products (X-Y, Y-Z, X-Z) using a slightly different testing procedure. Three t-tests were used for task time and satisfaction comparisons, chi-square tests were used for task completion and a non-parametric Mann-Whitney Test for error rate comparisons. Using this method of comparison increases the Family-Wise error rate. With a three group comparison using $\alpha = .05$, a p-value of less than .016 would need to be seen between any comparison in order to conclude one product showed performance significantly better or worse than chance.[2] For example, for the task completion rate, Product Z had a significantly higher ($p < .016$) task completion rate than Products X and Product X had a significantly higher task completion rate than Product Y ($p < .016$). The rankings in Table 6 (1-3) reflect the order of these differences. For Task time, however, while Product Y had a significantly longer task time than Product X, there was not a significant difference between Products X and Z. The interval graphs in Figure 2 also display the 95% Bonferroni confidence intervals for making multiple comparisons (notice the large overlap of confidence intervals between Products X and Z for task time).

Figure 2: Interval Plots of component usability metrics and SUM.*



*Mid-points are the mean values and the end-points represent the 95% Bonferroni confidence intervals. Note: A higher value is better for completion rates and satisfaction scores. A lower value is better for task time and errors.

[2] Since we did not have a hypothesis ahead of the competitive analysis on which product may perform better or worse than the others, a Bonferroni correction (alpha (.05) divided by the number of groups being compared (3)) was used. The Bonferroni confidence intervals in the Interval Graphs also include this correction. Without the correction, we are introducing an additional element of chance and the differences we observe at the $p < .05$ level will actually have a greater than 5% likelihood of being due to chance. This additional risk, called the Family Wise error rate, is 14%. Calculated as $1 - (1 - \alpha)^c$ where α is .05 and c is 3--the number of groups. For more information see [3] esp. pages 371 and 384.

3.2 Task Level Comparison

A task level comparison was performed to look for significant differences from any of the four metrics. Table 7 displays the means and standard deviations by task and product.

Table 7: Mean *Standard Deviation* for each component metric by task and product.*

	Completion			Time (seconds)			Satisfaction			Errors		
	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
Task												
1	92	100	95	**144 *85***	94 *26*	97 *33*	4.1 *.6*	4.0 *.5*	4.2 *.6*	0.9 *1.0*	**0.4 *0.6***	0.8 *0.7*
2	94	97	95	56 *56*	64 *47*	44 *34*	4.3 *.7*	4.1 *.9*	4.2 *.5*	0.4 *1.0*	0.2 *0.4*	0.1 *0.4*
3	86	**25**	80	207 *91*	**293 *116***	183 *88*	3.7 *1*	**2.6 *1***	4.0 *.9*	0.8 0.9	**2.3 *1.0***	0.6 *1.0*
4	65	61	90	142 *105*	136 *73*	**75 *68***	3.5 *1*	3.2 *1*	**4.1 *.5***	0.8 *1.1*	0.7 *0.6*	**0.2 *0.4***
5	61	78	**90**	135 *97*	140 *77*	92 *59*	3.8 *1*	3.5 *.8*	3.9 *.7*	1.2 *1.4*	0.7 *0.9*	0.6 *0.6*
6	90	97	100	177 *119*	194 *114*	144 *54*	3.9 *.9*	3.7 *.9*	**4.3 *.6***	1.2 *1.5*	1.3 *1.7*	0.4 *0.6*
7	94	**38**	90	274 *87*	**439 *133***	281 *74*	3.9 *.7*	**2.7 *1***	4.0 *.7*	1.7 *1.6*	**2.8 *1.6***	1.5 *1.0*
8	86	88	90	255 *102*	292 *113*	309 *119*	3.9 *.7*	3.7 *.8*	3.8 *.6*	1.0 *0.9*	1.5 *1.2*	1 *1.2.0*
9	69	53	**90**	189 *105*	182 *90*	134 *89*	3.3 *.9*	3.5 *.7*	3.8 *1*	1.4 *1.0*	1.9 *1.1*	**0.9 *0.6***
10	43	53	48	147 *85*	157 *85*	**210 *89***	3.4 *.9*	3.3 *.9*	3.0 *1*	1.2 *0.9*	0.9 *0.7*	1.3 *0.9*

* Bold values are significantly difference at the $p < .05$ level. A visual inspection of the confidence intervals was used to determine which product was significantly different than the other two.

Next, the SUM value was calculated from the average of the four standardized metrics. These standardized metrics are the values in the "% Equivalent" column from the Methods section above (Tables 3-5). Table 8 displays the SUM scores plus the standardized component metric values.

Table 8: SUM Scores plus standardized percentages for component metrics by task and product.*

	SUM %			Comp %			Time %			Sat %			Errors %		
	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
Task															
1	**70***	84	83	92	100	95	51*	91	86	55	53	64	84	93*	87
2	83	81	87	94	97	95	84	77	93	64	54	63	89	96	96
3	74	**39***	76	86	25*	80	75	46*	82	37	9*	48	91	75*	93
4	63	61	**84***	65	61	90	73	75	95*	32	22	56*	84	85	95*
5	60	63	**79***	61	78	90*	66	61	85	41	28	45	69	84	97
6	75	72	**88***	90	97	100	70	61	85	46	35	71*	93	93	98
7	81	**48***	83	94	38*	90	96	61*	98	44	10*	52	90	83*	91
8	77	71	72	86	88	90	81	71	64	45	35	39	94	91	94
9	61	56	**76***	69	53	90*	66	70	85	23	24	42	84	78	90*
10	44	45	**35***	43	53	48	38	26	9*	26	23	16	70	78	68

* **Bold** and **"*"**values show significant difference for SUM values. "*" only represent significance difference for the component metrics. Both are significant at the $p < .05$ level. All values are percentages. A higher value is better for all metrics.

Table 9 is a simplified view of the raw data displayed in Table 8 and shows which product had any significant difference in performance (better or worse). The letters in the cell indicate the Product that performed significantly better (+) or significantly worse (-) on the task and metric. For example, on Task 1, Product X had a significantly worse SUM score than Products Y and Z. Additionally, Product Y had significantly fewer errors on Task 1 than Products X and Z. There were no significant differences in any of the metrics for Tasks 2 and 8.

Table 9: Significant Difference for SUM and component metrics by Task.*

Task	SUM	Completion	Time	Satisfaction	Errors
1	**X-**		**X-**		**Y+**
2					
3	**Y-**	**Y-**	**Y-**	**Y-**	**Y-**
4	**Z+**		**Z+**	**Z+**	**Z+**
5	**Z+**	**Z+**			
6	**Z+**			**Z+**	
7	**Y-**	**Y-**	**Y-**	**Y-**	**Y-**
8					
9	**Z+**	**Z+**			**Z+**
10	**Z-**		**Z-**		

*Significant at the $p < .05$ level. The letters (X,Y or Z) in the cell indicate which product was significantly different for the task by the metric. The "+" indicates significantly better performance and the "–" indicates significantly worst performance.

For the significance tests, a One-way ANOVA was used for task time and satisfaction scores. A Chi-Square Test was used to compare completion rates and a non-parametric Kuskal-Wallis test was performed on the error data. The Bonferroni correction was not used since we were only looking for any significant difference between products by metric and task, not for all comparisons. The overlap in the confidence intervals was used to determine which product was significantly different than the others.

4 Discussion

4.1 Detecting Significant Difference at the Product Level

From a product usability level, using the boundaries of the confidence intervals on the graphs of the component metric data would most likely lead a business leader to conclude that Product Z is more usable that Products X and Y (Figure 2). This is especially the case for task completion and satisfaction that show little overlap in their confidence intervals, and not as clear for task time and error rates which have a lot of overlap. The interval graph of the SUM score however, paints an unequivocal picture of the ordering of Product Usability: Product Z is most usable, followed by Products X then Y. The large number of observations from each product (490, 318 and 209 respectively)[3] provides enough data to observe finer discriminations with the component metrics. As the number of observations decrease, either through running fewer subjects or fewer tasks, the ability to detect smaller differences will decrease. This is especially noticeable when making distinctions at the task level where the number of observations is much smaller than at the product level where all tasks are combined.

4.2 Detecting Significant Difference at the Task Level

Discerning the usability between products from a task level using only the component metrics wasn't as clear-cut as it was for the products as a whole (Tables 7-9). There are significant differences that show up for some metrics and not others. For example, in Task 1 Product Y has significantly fewer errors than the other products but this difference isn't significant for any of Product Y's other metrics on Task 1. Additionally, Product X shows a significantly worse task time, but no other significant differences in the component metrics for Task 1.

Most notably, Tasks, 1, 4, 6 and 10 exemplify the fact that important differences in usability don't always show up in the task completion rate (Table 9). Each of these tasks had significant difference in SUM scores, as well as significant differences in some combination of the component metrics but no significant difference in completion rates. What's more, the task completion rate becomes a less meaningful metric to discern differences as the completion rate approached 100% (as was the case in Tasks 1 and 6).[4]

What does it mean when users complete the task faster on one task but commit more errors on another? Is it acceptable to for users to complete more tasks but express less "satisfaction?" Is it really fair to classify a product or task as more usable if users are completing the tasks but taking an unacceptable amount of time? The confusion encountered when examining multiple measures, such as those in this competitive analysis, was also succinctly expressed elsewhere: "Which carries more weight: time to task completion or user satisfaction? I recognize this yearning for a single usability score."[5]

5 Conclusion

Using only one of the four component metrics to make decisions about usability is risky. Picking which metrics are more important for each task is also a risky method. With both strategies, the business leader or usability engineer is gambling that the important usability differences will show up in their chosen metric (often task completion).

Our recommended strategy is to start with the SUM results (product and task level) to get an initial assessment of the usability, then proceed to the component metrics to discover causes of a poor SUM score. If a business decision must be made using only one measure of usability, it's best to make that decision from a metric that summarizes the majority of variation in the four common usability metrics.

It is a difficult task for the usability engineer, the business leader and a potential customer of software, to determine which product or task is superior in usability while considering multiple measures on different scales. The advantage of a composite measure is in its ability to describe more than one attribute of the data. This benefit is especially helpful when comparing the usability of competing products or changes in the product over time.

---

[3] The total number of observations in the competitive analysis (1017) is different than total number of observations used to verify the correlations and Principal Component results as discussed in the introduction (1085). This discrepancy is a result of some tasks being left out of the competitive analysis as there was no overlap in functionality between all three products.

[4] To complicate matters, since task completion rates are modeled as a binomial proportion, more users are needed to distinguish small difference between the samples since the data is discrete. Continuous data such as task time and SUM scores can provide more discrimination with fewer users.

References

1. ANSI (2001). *Common industry format for usability test reports* (ANSI-NCITS 354-2001). Washington, DC: American National Standards Institute.
2. Breyfogle, F. (1999). *Implementing Six Sigma: Smarter Solutions Using Statistical Methods*. John Wiley and Sons.
3. Howell, David (2002) *Statistical Methods for Psychology* Fifth Edition. Duxbury Press, Pacifica Grove, CA
4. ISO. (1998). *Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability* (ISO 9241-11:1998(E)). Geneva, Switzerland: Author.
5. Jong, Steven (2000) *Musing on Metrics: Why Measure Usability?* Usability Interface, Vol. 7 No 2. http://www.stcsig.org/usability/newsletter/0010-metrics.html
6. McIver, J. P., & Carmines, E. G. (1981). *Unidimensional scaling*. Thousand Oaks, CA: Sage.
7. Nielsen, J. and Levy, J. (1994) Measuring Usability: Preference vs. Performance. *Communications of the ACM*, 37, p. 66-76
8. Nunnally, J. C. (1978). *Psychometric Theory*. New York, NY: McGraw-Hill.
9. Reichheld, Fredrick (2003). *The One Number you need to Grow* Harvard Business Review December 2003 (47-54)
10. Sauro, J & Kindlund E. (2005) "A Method to Standardize Usability Metrics into a Single Score." *in Proceedings of the Conference in Human Factors in Computing Systems 2005*
11. Sauro, J. (2005) "How long should a task take? Identifying Spec Limits for Task Times in Usability Tests." in *Proceedings of the 11th International Conference on Human-Computer Interaction*, Las Vegas USA. Also available at http://measuringusability.com/time_specs.htm
12. Sauro, J & Kindlund E. (2005) Making Sense of Usability Metrics: Usability and Six Sigma, in *Proceedings of the 14th Annual Conference of the Usability Professionals Association*, Montreal, PQ. Canada