# A Method to Standardize Usability Metrics Into a Single Score

**Jeff Sauro**
PeopleSoft, Inc.
Denver, Colorado USA
Jeff_Sauro@peoplesoft.com

**Erika Kindlund**
Intuit, Inc.
Mountain View, California USA
Erika_Kindlund@intuit.com

**ABSTRACT**
Current methods to represent system or task usability in a single metric do not include all the ANSI and ISO defined usability aspects: effectiveness, efficiency & satisfaction. We propose a method to simplify all the ANSI and ISO aspects of usability into a single, standardized and summated usability metric (SUM). In four data sets, totaling 1860 task observations, we show that these aspects of usability are correlated and equally weighted and present a quantitative model for usability. Using standardization techniques from Six Sigma, we propose a scalable process for standardizing disparate usability metrics and show how Principal Components Analysis can be used to establish appropriate weighting for a summated model.

SUM provides one continuous variable for summative usability evaluations that can be used in regression analysis, hypothesis testing and usability reporting.

**ACM Classification**
H5.2 [**Information Interfaces and Presentation**]: User Interfaces – Standardization; Benchmarking; Evaluation/Methodology.

**Keywords**
Usability; measurement; standardization; Principal Components Analysis; Six Sigma.

**INTRODUCTION**
In a summative usability evaluation, several metrics are available to the analyst for benchmarking the usability of a product. There is general agreement from the standards boards ANSI 2001[2] and ISO 9241 pt.11[18] as to what the dimensions of usability are (effectiveness, efficiency & satisfaction) and to a lesser extent which metrics are most commonly used to quantify those dimensions. Effectiveness includes measures for completion rates and errors, efficiency is measured from time on task and satisfaction is summarized using any of a number of standardized satisfaction questionnaires (either collected on a task-by-task basis or at the end of a test session) [2],[18].


*CHI 2005*, April 2–7, 2005, Portland, Oregon, USA.


### The Irony: Usability Metrics Need to be Easier to Use
As usability analysts encourage business leaders to track "usability" against other indicators of company performance—such as revenue growth, customer support expenditures or product abandonment rate—the various metrics we depend upon become clumsy and difficult to use. Each metric is measured on its own scale and yet each must be represented in the analysis and reporting process if we are to be true to the accepted industry definition of usability. Furthermore, differences in the scales make it difficult to compare the relative usability of different features or products.

This complexity in analysis and reporting makes usability data hard to digest. The analyst is challenged to present multiple usability metrics that clearly delineates usable and unusable aspects in a product without overwhelming business leaders or inadvertently promoting one metric over another.

To increase the meaningfulness and strategic influence of usability data, analysts need to be able to represent the entire construct of usability as a single dependent variable without sacrificing precision.

### Related Research
There have been attempts to derive a single measure for the construct of usability.

Babiker et al [3] derived a single metric for usability in hypertext systems using objective performance measures only. They found their metric correlated to subjective assessment measures but could not generalize their model to other systems.

Questionnaires such as the SUMI [22,23], PSSUQ[27], QUIS[7] and SUS[5] have users provide a subjective assessment of recently completed tasks or specific product issues and claim to derive a reliable and low-cost standardized measure of the overall usability or quality of use of a system. While the authors of these questionnaires do not necessarily intend for the questionnaires to act as a single

measure of usability (e.g. "QUIS was designed to assess users' subjective satisfaction with specific aspects of the human-computer interface" [7]), they are often used by practitioners as a way to measure usability with one number. Such usage is often not discouraged by the questionnaires' instructions (e.g. "SUMI is the only commercially available questionnaire for the assessment of the usability of software" [22] and "The SUS scale is a Likert scale and yields a single number representing a composite measure of the overall usability of the system [5]").

Cordes [8] and McGee [32, 33] used a method of magnitude estimation derived from methods in psychophysics as outlined by Stevens [45]. Specifically, McGee uses a geometric averaging procedure (UME) to standardize ratios of participants' subjective assessment ratings on tasks to derive a single score for task usability. His research identifies the potential for a standardized measure of usability to support usability comparisons across products, the same product over time, at lower levels of detail, and of tasks common to multiple products.

Lewis used a rank-based system when assessing competing products [25]. This approach creates a rank score comprised of both users' objective performance measures and subjective assessment, but the resulting metric only represents a relative comparison between like-products with similar tasks. It does not result in an absolute measure of usability that can be compared across products or different task-sets.

These methods provide helpful information to the analyst in making decisions about usability; however, one must question the ability of methods relying solely on objective or subjective measures to effectively describe the entire construct of usability in light of the guidance set by ISO 9241 and ANSI 354-2001 (a point also made by Dumas [9 esp p.1096]). Additionally, the reliance on relative ranking falls short of an absolute measure that can be freely compared as a standardized measure. Yet, the existence and usage of all these methods demonstrates the need to represent the complex construct of usability into a succinct and manageable form.

## BUILDING A QUANTITATIVE MODEL OF USABILITY

In an attempt to fully represent the entire construct of usability as well as creating a single usability metric we began with a high-level model of usability starting with the ISO/ANSI dimensions (effectiveness, efficiency & satisfaction). We used the following four metrics to represent these dimensions—task completion, error counts, task times and satisfaction scores (see Figure 1.)

To investigate the relationship among the metrics to properly build the model and a single score, we set up a data collection plan for four summative usability evaluations.

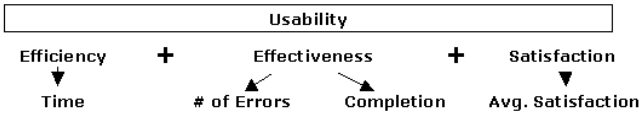


**Figure 1. Quantitative Model of Usability**

## METHOD

Four summative usability tests were conducted to collect the common metrics as described above (task completion, error counts, task times and satisfaction scores) as well as several other metrics as suggested in Dumas and Redish [11], and Nielsen [39]. For measuring satisfaction we created a questionnaire containing semantic distance scales with five points, similar to the ASQ created by Lewis [26] (see Table 5 below). The questionnaire included questions on task experience, ease of task, time on task, and overall task satisfaction. The questionnaires were administered immediately after each task to improve accuracy [16]. The four usability tests were conducted in a controlled usability lab setting over a two-year period. Participants were asked to complete the tasks to the best of their ability and the administrator only intervened when the participant indicated they were done or gave up. At the end of the test session, "post-test" satisfaction questions similar to those in SUMI and SUS that asked about overall product usability were given to users.

The applications tested were all from the financial and accounting industry. Three were Windows-based and one application was web-based. One application was tested twice one year apart. Each test used different test administrators (one administered two tests) in five geographic locations within the US. Data was collected from 129 total participants completing a total of 57 tasks. Participants varied in their application experience, gender, and industries.

## RESULTS

### Examining the Relationships between the Metrics

To attempt to combine the metrics into a single usability score we examined the relationship among the four primary variables for each task observation. We generated a correlation matrix with all four variables from all four data sets plus a combined data set containing data from all tests.

As can be seen in the lower right cell of Table 1, the Pearson Product Moment correlation coefficients between satisfaction and task completion are consistent with prior correlation analyses (that is, displaying moderate and significant correlations between .3 and .5) [26, 29]. What's more, the positive correlation between subjective measures (satisfaction) and objective measures (time, errors and completion) are also consistent with Nielsen's 1994 meta-analysis [38] (although the subjective measures were preferences instead of satisfaction in that study).

| | Time | Errors | Satisfaction |
|---|---|---|---|
| **Errors** | | | |
| A | .490 | | |
| B | .594 | | |
| C'03 | .578 | | |
| C'04 | .523 | | |
| Combined | .517 | | |
| **Satisfaction** | | | |
| A | -.379 | -.396 | |
| B | -.454 | -.449 | |
| C'03 | -.512 | -.403 | |
| C'04 | -.464 | -.286 | |
| Combined | -.478 | -.348 | |
| **Completion** | | | |
| A | -.145 | -.428 | .369 |
| B | -.403 | -.492 | .410 |
| C'03 | -.302 | -.380 | .503 |
| C'04 | -.251 | -.380 | .433 |
| Combined | -.268 | -.384 | .454 |

**Table 1: Correlation Matrix of Four Usability Metrics by Task Observation for Five Data Sets (All correlations $p$ <.001)**

**Key for Table 1**

| Data Set | Observations | # of Participants |
|---|---|---|
| Product A | 294 | 21 |
| Product B | 144 | 11 |
| Product C '03 | 644 | 48 |
| Product C '04 | 778 | 49 |
| Combined Data | 1860 | 129 |

Frøkjær et al [12] earlier has made the case for including all aspects (effectiveness, efficiency and satisfaction) when measuring the usability of a system since it was found that these aspects did not always correlate in the data they reviewed. We agree with Frøkjær et al's conclusion to measure all aspects of usability, however, not because they do not correlate with each other (our data clearly shows the opposite), but because each measure adds additional information not contained in the other measures.

The average values from the post-test satisfaction questions also showed low to moderate and significant correlations with average task performance by user (r = .22 to .33 $p$ <.10). The correlations were not as strong or as significant as the post-task satisfaction questions but still showed a similar relationship. We used the values from the post-task questions since the focus of our analysis was at task-level usability and this provided us with the same number of observations for all four variables.

### Summarizing Variables using Principal Components Analysis

When variables that are ostensibly measuring the same event correlate with each other, there is redundant information making analysis more complicated. Principal Components Analysis (PCA) [20] is a statistical technique that is commonly used in such situations. PCA linearly transforms an original set of variables into a smaller set of uncorrelated variables that represents most of the information in the original set of variables. Its goal is to reduce the dimensionality of the original data set. PCA is not to be confused with Factor Analysis, another common multivariate technique that can use a method of Principal Components. Factor Analysis aims at revealing the underlying structure of the data from many variables, whereas the aim of PCA is to explain the maximum amount of variance with the fewest number of components (See [20] esp. Ch 7). A smaller set of uncorrelated variables can be much easier to understand and use in further analyses than a larger set of correlated variables.

The variables used in a PCA need not be normally distributed or all continuous. The variables can be a mixture of continuous, ordinal and binary variables [20 esp. p. 339]. This flexibility makes PCA an especially helpful procedure in interpreting usability data that takes the form of continuous (time), ordinal (satisfaction) and binary (completion). The technique has been used to summarize behavioral data in the social sciences [36],[10] and [20].

Using the output of a Principle Components Analysis we: (a) build a better model of usability by minimizing the random error from any one measure (b) remove redundant data from the overlapping variables (e.g. if information contained in errors can be accounted for by time or satisfaction) and (c) uncover which, if any of the four original variables, weigh more heavily in the model. For example, task completion or satisfaction may account for more variance than time or errors.

### PCA Results and Output

The first step in interpreting the results of a Principal Components Analysis is to determine which components to retain. There are several methods and none are definitive, with each method requiring some level of judgment. Some of the more commonly used methods include:

1. **Kaiser's Rule:** Only retain principal components (PCs) with eigenvalues greater than 1. [21]. Jolliffe recommends .7 as more rigid cutoff [19].
2. **Scree Plot Test:** Stop retaining components at the point in a plot of the eigenvalues when the line levels off more or less horizontally similar to a pile of rocks at the bottom of a hill [6].
3. **Cumulative Variance:** Stop retaining when the cumulative variance of the PC's reach a certain predetermined level. This level fluctuates depending on the goal of the analysis. At a minimum the majority (greater than 50%) should be accounted for by the PCs and ideally 70%-90% [20].

The results of the analysis revealed similar results for all five data sets. We retained the first PC based on it meeting all three criteria listed above. Only the first PC contained eignervalues greater than 1 in all data sets (see Table 2 in row "PC 1"). As per method 2, the Scree plots of the eigenvalues also indicate only retaining the first PC (see Figure 2). The

cumulative variance from the first PC meets the minimum requirements (>50%) from method 3 (see Table 3).
After determining the number of components to be retained, the next step is to identify the construct that the retained PCs measure and assess which variables account for more of the variance.

As can be seen in Table 4, all four variables are significant (having coefficients greater than .3 [14]). Since all the variables are showing significant coefficients it indicates that each variable adds new information not contained in the other variables. That is, if we saw the coefficients for errors consistently falling below .3 we would conclude that errors are not adding a sufficient amount of new information to the combined model. An interpretation of the coefficients would read that as errors and time decrease, completion and satisfaction increase. This relationship is generally regarded as the construct of "usability."

Since all four variables have roughly equal coefficients on the first principal component across all five data sets we concluded that all four variables account for the same amount of variance—they are equally weighted. This result is consistent with Nunnally [40 esp. p. 297] who found that unweighted measures generally correlated so highly with weighted measures that it is seldom worth the effort in determining weights.

| | A | B | C'04 | C'03 | Combined |
|---|---|---|---|---|---|
| **PC 1** | 2.1171 | 2.4038 | 2.1712 | 2.3390 | 2.2260 |
| **PC 2** | 0.8575 | 0.6071 | 0.7943 | 0.7681 | 0.7667 |
| **PC 3** | 0.6248 | 0.5713 | 0.6764 | 0.5423 | 0.6260 |
| **PC 4** | 0.4005 | 0.4177 | 0.3582 | 0.3506 | 0.3813 |

**Table 2. Eigenvalues for Principal Components (PC) by Data Set (Only the first component in each data set has an eigenvalue >1)**

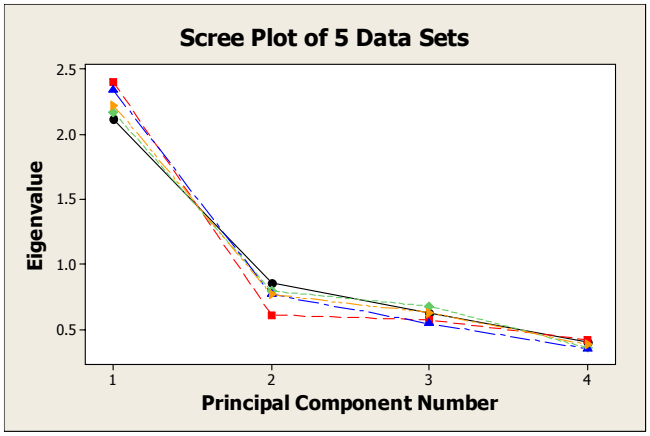


**Figure 2. Scree Plot of the Eigenvalues from all 5 Data Sets showing break in the 2nd PC indicating only 1 PC should be retained.**

| | A | B | C'04 | C'03 | Combined |
|---|---|---|---|---|---|
| **PC 1** | 0.529 | 0.601 | 0.543 | 0.585 | 0.556 |
| **PC 2** | 0.214 | 0.152 | 0.199 | 0.192 | 0.192 |
| **PC 3** | 0.156 | 0.143 | 0.169 | 0.136 | 0.156 |
| **PC 4** | 0.100 | 0.104 | 0.090 | 0.088 | 0.095 |

**Table 3. Proportion of Variance Accounted for Each Principle Component (PC) by Data Set (PC 1 accounts for a sufficient amount of the variance)**

| | Data Sets | | | | |
|---|---|---|---|---|---|
| | A | B | C'04 | C'03 | Combined |
| **Variance** | 52.9% | 60.1% | 54.3% | 58.5% | 55.6% |
| **Errors** | -.561 | -.526 | -.507 | -.508 | -.508 |
| **Time** | -.479 | -.524 | -.525 | -.517 | -.515 |
| **Completion** | .445 | .478 | .463 | .453 | .461 |
| **Satisfaction** | .508 | .470 | .503 | .519 | .515 |

**Table 4. Coefficients of the 1st Component called "Usability"**

### Components 2, 3 and 4

We examined the relationship between the variables in the remaining three principal components in all data sets and did not find a pattern that was interpretable. While the second component contained as much as 21% of the variance (see Table 3, row "PC 2"), there was no consistent discernable pattern. PC's 3 and 4 accounted for very little variance and could not be interpreted, as is common with the last PCs [20].

Due to our desire for parsimony, the aforementioned inconsistencies and the first PC meeting the requirements in methods 1-3, we retained only the first principal component.

### Using the PC Scores as a Surrogate Variable: A Summated Usability Score

Next we stored values (called scores) of the first PC and used these values as surrogate variables. PC scores are created by multiplying the variable coefficients by the raw variable data in standardized form and summing the products [20]. This creates one surrogate value instead of four. The surrogate variable is a composite of the four raw variables and accounts for between 52% and 60% of the variance. This variable represents the best mathematical combination of all four variables. It can be thought of as a "usability score" and can be used in the same way as any of the four variables can.

If the usability analyst has access to statistical software to run a PCA and store the scores from the first component, then he or she can use those scores for regression analysis, hypothesis testing and drawing conclusions in experimental analyses.

### Limitations of Using PC Scores as Surrogate Variables

The major drawback to using the stored scores is that they are dependent on the raw data used for that study and therefore cannot be compared to other component scores from other data sets. To compare scores across tests a summated scale needs to be created that duplicates the relationship built from the PCA [14]. Doing this requires standardizing all variables and then multiplying them times the coefficients from the first PC. Since the coefficients were consistently equal across the data sets, taking the arithmetic mean of the four standardized variables (or multiplying each by .25 and summing them) will provide similar values as the PCA scores. This is similar to the method succinctly described by Martin and Bateson [31]:

> *Measures that are to be combined usually need to be standardized so that they have the same mean and variation. One-way is to calculate for each raw value its z score: the score for that [observation] minus the mean score for the sample divided by the standard deviation. Scores standardized in this way have a mean of zero and a standard deviation of 1. The composite score for an individual is then the average of the z scores of the separate measures. This procedure gives the same statistical weight to each measure. If different weights are to be given to the separate measures, this is best done explicitly by multiplying the z score of each measure by an amount that can be specified; for instance, by its loading on a principal component, obtained by principal component analysis. p124*

The goal then becomes standardizing the four variables (time, satisfaction, completion and errors).

## STANDARDIZING USABILITY METRICS

To standardize each of the usability metrics we created a z-score type value or z-equivalent. For the continuous data (time and average satisfaction), we subtracted the mean value from a specification limit and divided by the standard deviation. For discrete data (completion rates and errors) we divided the unacceptable conditions (defects) by all opportunities for defects. This method of standardization was adapted from the process sigma metric used in Six Sigma [4],[17], [43]. See Sauro & Kindlund [44] for a more detailed discussion on how to standardize these metrics from raw usability data.

### Standardizing Task Completion

We can assume that all users want to successfully complete tasks, so a defect in task completion can be identified as an instance of a user failing a task. An opportunity for a defect in task completion is simply each instance of a user attempting a task. Therefore, we standardized task completion as the ratio of failed tasks to attempted tasks. This proportion of defects per opportunities has a corresponding z-equivalent that can be looked up in a standard normal table. For example, a task completion rate of 80% would have the z-equivalent of .841.

### Standardizing Error Rates

Unlike the calculation for task completion, it is insufficient to define "error opportunities" as simply each instance of a user in the sample attempting a task. This is because not all tasks are equal when it comes to error potential and users can commit more than one error per task (unlike task completion where they can only fail once per task). Complex tasks with many required components for task success have a greater potential for error then less complex tasks [41]. Our standardization process needs to account for this variation in error potential when trying to calculate the error probability.

Therefore, we defined a task's "error opportunities" as the number of sub-tasks that a user must conform to in order to complete a task error-free. This method is similar to calculating the Human Error Probability (HEP) as described in [41] and [15]. "The general approach for [determining HEP] is to divide human behavior in a system into small behavioral units, find data for these subdivisions and then recombine them to estimate the error probabilities for the task [41]."

Here is an example of how we defined a task's opportunities in terms of its sub-tasks:

**Example Task:** Add a new customer record to the Customer List

- **Opportunity 1**: Locate access point for adding a new customer and launch data form
- **Opportunity 2**: Enter new customer record ID information
- **Opportunity 3**: Enter account opening balance information correctly
- **Opportunity 4**: Enter customer address information
- **Opportunity 5**: Enter customer contact information
- **Opportunity 6**: Submit record successfully

While there are 6 opportunities for the user to make errors, there can be multiple ways an error can be committed. It's important to note that identifying opportunities does not mean identifying ideal paths through the software. Users may take many paths or choose many directions to accomplish certain tasks. If certain required operations are not completed, it's an error regardless of how the user arrived at the screen. For example, Opportunity #1 can have the following error instances associated with it:

- User can't find access point
- User launches an existing customer record instead of adding a new one
- User launches a new vendor record instead of a new customer record

Each error instance is unique, yet all are associated with the more general "opportunity" to make an error in this component of the task. Once the task's error opportunities

have been identified, the z-equivalent can be calculated by dividing the total number of errors by the error opportunities. This proportion can be approximated using the standard normal deviate.

### Standardizing Satisfaction Scores

As described in the Methods section, we used a post task questionnaire containing 5-point semantic distance scales with the end points labeled (e.g. 5:Very Easy to 1:Very Difficult). 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 5) .The three questions had high internal-reliability (coefficient alpha .92, .91, .91 .89 for the four data sets). 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 [34 esp p.15], [27], [40] and [13].

To standardize the composite score we looked to the literature for a logical specification limit. Prior research across numerous usability studies suggests that systems with "good-usability" typically have a mean rating of 4 on a 1-5 scale and 5.6 on a 1-7 scale [38]. Therefore we set the specification limit to 4. To arrive at a standardized z-equivalent for composite satisfaction we subtracted the average rating of a user's satisfaction score from 4 and divided by the standard deviation.

While the specification limits of 4 (5-point scales) and 5.6 (7-point scales) are good guideposts for setting specification limits they should be used as starting points.

| | | | | |
|---|---|---|---|---|
| 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 |

**Table 5. Post-Task Satisfaction Questions used in building Composite Satisfaction**

Analysts should always investigate data for the specific domain that would either confirm these values as appropriate spec limits or specify slightly higher or lower values.

### Standardizing Task Times

Identifying ideal task times presents an interesting challenge: how long is too long for any given task? When comparing task times between products, a simple T-Test of the means will identify significant differences. For looking at only one set of times, the point at which a task takes too long is not as easy to define. It is not indefinable, just difficult to define in an absolute sense without some arbitrariness. Lewis [24], Whiteside, Bennett and Holtzblatt [46] and Sauro [42] offer recommendations on identifying specification limits for task times.

Once the ideal task time has been set for each task, standardizing the task time involves subtracting the raw task time from the specification limit and dividing by the standard deviation to arrive at the z-equivalent.

### Creating a Single, Standardized and Summated Usability Metric: SUM

We created a single, standardized and summated usability metric for each task by averaging together the four standardized values based on the equal weighting of the coefficients from the Principal Components Analysis. To be sure our method of standardization was properly reflecting the relationship built from the PCA, we regressed the scores from the 1st PC with the average of our four standardized metrics for each data set (see Figure 3 for an example from data set C'04).

As can be seen from the Fitted Line Plot in Figure 3, there is a very strong positive correlation between the scores calculated from the PCA and from the method of standardization [(R-Sq adj (82.1%) $p < .001$)].

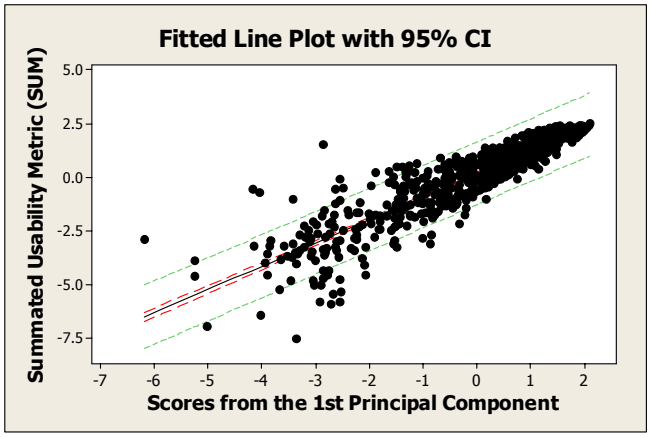


**Figure 3. Regression Plot of PCA Score on the Standardized Summated Usability Metric for Product C04 (776 Observations) (R-Sq adj (82.1%) $p < .001$).**

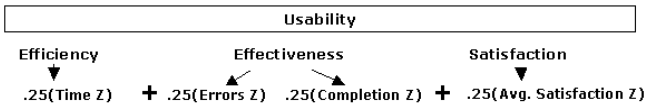


**Figure 4. A Weighted Quantitative Model of Usability**

This strong relationship suggests that using the average of the four standardized metrics will adequately mimic the relationship built from the Principal Components Analysis. This standardized and summated usability metric can now be used for analysis as well as for comparisons across tasks and studies. Our Quantitative Model of Usability (see Figure 4) now reflects the equal weighting of the standardized component metrics to summarize the construct of usability.

## DISCUSSION

### Scalability and Limitations

Four data sets and 1860 observations provide a starting point for further investigating this relationship between usability metrics. It is encouraging that similar results were obtained under different testing conditions, with different products and using different test administrators with slightly different testing protocols. Testing a greater variety of products with a broad spectrum of users will provide more insight into the validity of this model and approach. Future analyses are necessary to provide an indication of how versatile this model is in different domains and with different interfaces (both hardware and software).

As stated by Molich, et. al. [35], the effectiveness of a usability test is dependent on the chosen tasks, the methodology, and the persons in charge of the test. We acknowledge that the reliability of any metrics procured from a summative evaluation can be equally dependant on these factors. However, having a model for deriving a standard measure is also a powerful tool to evaluate differences in testing procedures.

### Are We Measuring Usability?

Our method summarizes the majority of variance in four metrics commonly used to assess the usability of a product in a summative evaluation. Whether these metrics properly quantify "usability" is a much larger discussion and we do not claim to be definitively measuring the construct of usability.

A summated usability metric is only as good as its underlying component metrics and to the extent that ISO and ANSI have properly identified those is certainly worth discussion. Others might add more metrics to a summative model, such as measures for learnability or memorability [36], [1]. Still others might argue for fewer measures for the sake of expediency or to remove subjectivity. For example, identifying errors and error opportunities is both time consuming and arguably the most subjectively built metric—not all analysts will agree on what constitutes an error or error opportunity. Errors are also not always included in models of usability [2],[30].

### Reducing the Number of Variables in the Model

There are strong opinions both for and against including errors in a summative model. We excluded errors from our PCA analysis and found that the 1st PC can still summarize the majority of variance in the three remaining variables. The three-variable model also had roughly equally weighted variables (although satisfaction weighed slightly more heavily in 3 of the four data sets) - see Table 6.

Error analysis plays a crucial role in formative evaluations when the goal is to uncover usability problems in an interface. In our data sets, users performed a task successfully, quickly and reported a high level of satisfaction yet committed some undesirable errors. Only the error measurement reflected this "unusable" aspect of the task. Errors have a coefficient as strong as any of the other variables in the first principal component (see Table 4) indicating they provide additional information not contained in the other three variables. For these reasons we find their inclusion in summative evaluations worth the effort.

### Increasing the Number of Variables in the Model

We also examined the relationship when including two additional metrics—help access and click counts. The two metrics have significant and moderate to strong correlations with the existing four metrics (r =.2 to .6 $p$ <.01). We included each variable in the PCA for the respective data sets to see how their inclusion affected the variable weights (See Tables 7 and 8).

In data set B, the addition of click counts moderately affected the coefficients and slightly reduced the variance. Click counts correlated highly with task time and errors (r = .670 and .589 $p$ <.001 respectively).

In data set C'04, Help was accessed in 72 of the 778 observations or about 10% of the time. Its inclusion also slightly changed the coefficients and brought the amount of variance down below 50% for the 1st PC (See Table 8).

| | A | B | C'04 | C'03 | Combined |
|---|---|---|---|---|---|
| **Variance** | 53.6% | 62.3% | 59.1% | 58.5% | 60.2% |
| **Time** | -.538 | -.594 | -.556 | -.551 | -.555 |
| **Completion** | .527 | .571 | .537 | .546 | .540 |
| **Satisfaction** | .657 | .567 | .635 | .630 | .633 |

**Table 6. Variance and Coefficients of the 1st Component, Errors excluded**

| Product B | PC 1 | PC 2 |
|---|---|---|
| **Variance** | 57.9% | 16.6% |
| **Errors** | -.484 | .033 |
| **Time** | -.497 | .166 |
| **Completion** | .372 | .776 |
| **Satisfaction** | .413 | .182 |
| **Click Count** | -.458 | .580 |

**Table 7. Variance and Coefficients of the 1st and 2nd PCs Click count added for Product B Only**

| Product C'04 | PC 1 | PC 2 |
|---|---|---|
| **Variance** | 47.9% | 16.8% |
| **Errors** | -.457 | .204 |
| **Time** | -.504 | -.261 |
| **Completion** | .422 | -.581 |
| **Satisfaction** | .470 | -.138 |
| **Help Access** | -.372 | -.730 |

**Table 8. Variance and Coefficients of the 1st and 2nd PCs Help Access added for Product C'04 Only**

While both these variables were helpful for analyses in their raw form, we decided against including them in the model since both are infrequently collected in summative evaluations at our organizations and therefore would impede cross-product comparisons.

If the goals of a summative evaluation require certain metrics to be evaluated, then this method of combining standardized metrics can still be used. The analyst should check the correlation of the metrics and run a PCA to assess the coefficients for weighting and amount of variance explained. All things being equal, it's better to include more variables than less in a summative metric. The point of diminishing returns occurs when variables added reduce the amount of variance accounted for by one PC to below 50%. This did not occur with the addition of click counts in data set B (57.9% variance) but did with the addition of Help Access in Data Set C'04 (47.9%).

The major drawback with adding or subtracting variables would be that a score created with 3 variables cannot be compared to a score created with 4 or 5 variables. Adopting a standard that captures the majority of the variance based on the most universal metrics is recommended.

## CONCLUSION

A single, standardized and summated usability metric (SUM) cannot and should not take the place of diagnostic qualitative usability improvements typically found in formative evaluations. When a summative evaluation is used to quantitatively assess the "before and after" impact of design changes, the advantage of one score is in its ability to summarize the majority of variance in four integral summative usability measures. SUM has two additional advantages. First it provides one continuous variable that can be used in regression analysis, hypothesis testing and in the same ways existing metrics are used to report usability. Second, a single metric based on logical specification limits provides an idea of how usable a task or product is without having to reference historical data. This score can then be used to report against other key business metrics.

SUM can never replace all the information inherent in the component metrics, but like a FICO score, an IQ score or even the Richter scale, the ability to provide high-level summary information about a complex construct with one number should prove helpful for informing and making decisions about usability.

Data from four summative evaluations indicates that our model provides a versatile method that can be used to develop a single, standardized and summated score for analyzing and reporting usability metrics.

## ACKNOWLEDGEMENTS

The authors would like to thank Intuit, Inc. for providing the facilities and opportunity to conduct research within usability testing. We further thank Grace Pariante, Shara Barnett and Jen Moore for their support collecting data used in our analysis, Jim Lewis, Lynda Finn, Rolf Molich, Wayne Gray and Kaaren Hanson for reviewing our research and providing feedback on previous versions of this paper.

## REFERENCES

1. Abran, A., Surya, W., Khelifi, A., Rilling, J., Seffah, A., Robert, F. (2003). Consolidating the ISO Usability Models. Paper presented at *11th annual International Software Quality Management Conference*.
2. ANSI (2001). *Common industry format for usability test reports* (ANSI-NCITS 354-2001). Washington, DC: American National Standards Institute.
3. Babiker, E.M., Fujihara, H., Boyle, Craig. D. B. (1991). A metric for hypertext usability. In *Proc. 11th Annual International Conference on Systems documentation,* (pp.95-104*).* ACM Press.
4. Breyfogle, F. (1999). *Implementing Six Sigma: Smarter Solutions Using Statistical Methods*. John Wiley and Sons.
5. Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. Jordan, B. Thomas, and B. Weerdmeester (Eds.), *Usability Evaluation in Industry* (pp.189-194*)*. London: Taylor and Francis. See also http://www.cee.hw.ac.uk/~ph/sus.html
6. Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245– 276.
7. Chin, J. P., Diehl, V. A., and Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. *Proc. CHI '88*. (pp. 213-218). Washington, D.C.: ACM Press. See also http://www.lap.umd.edu/QUIS/index.html
8. Cordes, R. E (1984). Application of Magnitude Estimation for Evaluating Software Ease of Use. In Gavriel Salvendy (Ed.) *First USA-Japan Conference on Human Computer Interaction*, Amsterdam: Elsevier Science Publishers.
9. Dumas, J. S. (2003). User-based evaluations. In J. A. Jacko and A. Sears (Eds.), *The Human-Computer Interaction Handbook* (pp. 1093-1117). Mahwah, NJ: Lawrence Erlbaum.
10. Dunteman, George H, (1989) Principal Components Analysis. In Sage University Papers Series *Quantitative Applications in the Social Sciences ; No. 07-069* Newbury Park Sage Publications, Inc.
11. Dumas, J., and Redish, J. C. (1999). *A practical guide to usability testing*. Portland, OR: Intellect.
12. Frøkjær, E., Hertzum, M., and Hornbæk, K. (2000) Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In *Proc. CHI 2000*, (pp.345-352). Washington, D.C.: ACM Press.
13. Gliem, J. and Gliem, R. (2003). Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales. In *2003 Midwest Research to Practice Conference in Adult, Continuing and Community Education.* Columbus, OH.
14. Hair, Anderson, Tatham, Black (1998) *Multivariate Data Analysis Fifth Edition*. NJ: Prentice Hall.

15. Hagen, E. (Ed.) (1976). Human Reliability Analysis, control and instrumentation. *Nuclear Safety*. 17(3), 315-326.
16. Hassenzahl, M. Sandweg, N. (2004). From Mental Effort to Perceived Usability: Transforming Experiences into Summary Assessments. In the *Extended Abstracts of the 2004 conference on Human Factors and Computing Systems* (pp 1283-1286).
17. Harry, M. J (1987). The Nature of Six Sigma Quality. *Technical Report, Government Electronics Group*, Motorola Inc. Scottsdale, AZ.
18. 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.
19. Jolliffe, I. T. (1972). Discarding variables in a principal component analysis 1: Artificial data. *Applied Statistics*, 21, 160– 173.
20. Jolliffe, Ian T.(2002). *Principal Component Analysis*. Secaucus, NJ, USA: Springer-Verlag.
21. Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational Psychology Measurement*, 20, 141– 151.
22. Kirakowski, J. (1996). The Software Usability Measurement Inventory: Background and usage. In P. Jordan, B. Thomas, and B. Weerdmeester (Eds.), *Usability Evaluation in Industry* (pp. 169-178). London, UK: Taylor and Francis. (Also, see http://www.ucc.ie/hfrg/questionnaires/sumi/index.html )
23. Kirakowski, J., and Corbett, M. (1993). SUMI: The Software Usability Measurement Inventory. *British Journal of Educational Technology*, *24*, 210-212.
24. Lewis, J. R (1982) "Testing Small System Customer Setup" in *Proceedings of the Human Factors Society 26th Annual Meeting* p. 718-720
25. Lewis, J (1991) A Rank-Based Method for the Usability Comparison of Competing Products. In *Proceedings of the Human Factors and Ergonomics Society 35th Annual Meeting San Francisco California* (pp1312-1316).
26. Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ. *SIGCHI Bulletin*, *23*, 78-81.
27. Lewis, J. R. (1992). Psychometric evaluation of the Post-Study System Usability Questionnaire: The PSSUQ. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 1259-1263). Atlanta, GA: Human Factors Society.
28. Lewis, J. R. (1993). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use (Tech. Report 54.786). Boca Raton, FL: IBM Corp. http://drjim.0catch.com/usabqtr.pdf
29. Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7, 57-78.
30. Lewis, J.R. (In Press). *Handbook of Human Factors and Ergonomics 3rd Edition*. Gavriel Salvendy Editor. John Wiley & Sons.
31. Martin,P and Bateson, P (1993) *Measuring Behaviour.* (2nd Edition) Cambridge [England]; New York, NY, USA : Cambridge University Press.
32. McGee, M. (2003). Usability magnitude estimation. *Proc. HFES, 47th Annual Meeting*, (691--695).
33. McGee, M (2004). Master usability scaling: magnitude estimation and master scaling applied to usability measurement. In *Proc. CHI 2004,* (pp 335 - 342). Washington, D.C.: ACM Press.
34. McIver, J. P., & Carmines, E. G. (1981). *Unidimensional scaling*. Thousand Oaks, CA: Sage.
35. Molich, R., Ede, M., Kaasgaard, K., and Karyukin, B (2004). Comparative Usability Evaluation. *Behaviour & Information Technology*, 23(1), 65-74.
36. Morrison, D. F. (1976). *Multivariate statistical methods* (2nd ed.). New York: McGraw-Hill
37. Nielsen, J (1994) *Usability Engineering*. San Francisco: Morgan Kaufman.
38. Nielsen, J. and Levy, J. (1994) Measuring Usability: Preference vs. Performance. *Communications of the ACM*, 37, p. 66-76
39. Nielsen, J. (1997). Usability testing. In G. Salvendy (Ed.), *The Handbook of Human Factors and Ergonomics,* (2rd Edition). John Wiley & Sons.
40. Nunnally, J. C. (1978). *Psychometric Theory*. New York, NY: McGraw-Hill.
41. Park, Kyung S. (1997). Human Error. In Gavriel Salvendy (Ed.), *The Handbook of Human Factors and Ergonomics,* (2rd Edition). John Wiley & Sons.
*42*. Sauro, J. (2004) How long should a task take? Identifying Spec Limits for Task Times in Usability Tests. *Retrieved September 13, 2004, from Measuring Usability Web site : http://measuringusability.com/time_specs.htm*
*43*. Sauro, J. (2004) How Do You Calculate a Z-Score? *Retrieved September 13, 2004, from Measuring Usability Web site: http://measuringusability.com/z_calc.htm*
44. Sauro, J & Kindlund E. (In Press) Making Sense of Usability Metrics: Usability and Six Sigma, in *Proceedings of the 14th Annual Conference of the Usability Professionals Association*, Montreal, Canada
45. Stevens, S.S. (1975). *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. New York: John Wiley.
46. Whiteside, J., Bennett, J. and Holtzblatt, K. (1988) "Usability Engineering: Our Experience and Evolution" in *The Handbook of Human Computer Interaction* Elsevier Science Publishers, Amsterdam, pp 791-817