Explorations on Single Usability Metrics

Michael Van Waardhuizen
Microsoft
Redmond, WA 98052, USA
micvan@microsoft.com

Jennifer McLean-Oliver
Microsoft
Redmond, WA 98052, USA
jennimc@microsoft.com

Nancy Perry
Microsoft
Redmond, WA 98052, USA
nancyp@microsoft.com

Joe Munko
Microsoft
Redmond, WA 98052, USA
jmunko@microsoft.com

ABSTRACT

A long-term summative evaluation program was undertaken at Microsoft. This program focused on generating a Single Usability Metric (SUM) score across products over time but encountered a number of issues including error rate reliability, challenges establishing objective time-on-task targets, and scale anchoring. These issues contributed to making SUM difficult to communicate, prompting exploration of an alternative single usability metric using simple thresholds developed from anchor text and inter-metric correlations.

CCS CONCEPTS

• **Human-centered computing~**Usability testing

KEYWORDS

Usability; usability measurement; single usability metrics

CHI'19 Extended Abstracts, May 4-9, 2019, Glasgow, Scotland, UK.

ACM ISBN 978-1-4503-5971-9/19/05. DOI: https://doi.org/10.1145/3290607.3299062

ACM Reference format:

Michael Van Waardhuizen, Jennifer McLean-Oliver, Nancy Perry, and Joe Munko. 2019. Explorations on Single Usability Metrics. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'19 Extended Abstracts), May 4–9, 2019, Glasgow, Scotland, UK.* ACM, New York, NY, USA. 8 pages. https://doi.org/10.1145/3290607.3299062

1 INTRODUCTION

In 2015, a small research team was set up within Microsoft with the express purpose of running summative evaluation usability studies (benchmarks.) Prior to this, benchmarks were typically conducted infrequently, leading to variable outcomes and difficulty comparing products and scenarios across the company. With the goal of enabling better comparisons, this research team was chartered to develop a set of best practices for conducting usability measurements, including the selection and computation of metrics to enable better communication of usability with product stakeholders. The metrics were initially selected to enable the calculation of the Single Usability Metric (SUM) proposed by Jeff Sauro and Erika Kindlund in 2005 [4][5].

Over the course of 2.5 years, 36 studies were completed, collecting data for almost 500 tasks across 13 platforms and 800 users (combined total of more than 9000 observations.) Along the way, many lessons were learned that required an evolution of our metrics away from the initial proposal. This paper presents some of the lessons learned with SUM, and the subsequent steps taken to measure and communicate usability effectively.

2 METHOD

All of the benchmark evaluations run over this period used the same base set of metrics, derived from Sauro & Kindlund's SUM. These metrics included:

- Task Completion (Success)
- Time on Task
- Errors
- Perceived Time (5 pt Likert scale)
- Perceived Difficulty (5 pt Likert scale)
- Satisfaction (5 pt Likert scale)

The above metrics were collected across in-person/remote, moderated/unmoderated, between group/within group, and across a variety of device platforms. They were used and computed into a SUM per the guidance in the original paper. Additional task metrics as well as summary metrics such as SUS and NPS were collected but are not included for this analysis and discussion.

A single usability metric was pursued for the goal of enabling better communication with stakeholders. It can provide a "temperature gauge" number to directly answer the question "How usable is it?" A single metric should incorporate both subjective and objective measurements and so provide a more rigorous and balanced look at usability than just one or the other. It can also help avoid confusion among multiple metrics, especially when trying to judge whether a design is better than a previous one or a competitor.

Broadly speaking, SUM is calculated as follows:

- Subjective measures are averaged, adjusted by a "spec limit", and converted to z-values by dividing by the standard deviation. The recommended spec limit is 4, on a 5 pt Likert scale.
- Time on Task is first transformed into a normal distribution using natural logarithms, then adjusted by a spec limit and z-values are calculated. The spec limit recommended is an expert time multiplied by 3.
- Errors are measured and divided by the number of error opportunities to get an error rate.
- Completion rate and error rate are converted to z equivalents as well to be on the same basis.

The SUM is then computed by averaging the four components and converting back into a percentile.

3 FINDINGS AND DISCUSSION

Our experience with attempting to use SUM uncovered several challenges, either in calculation or in communication.

3.1 Error Counting

In their original proposal, Sauro and Kindlund recommend errors as one of the four primary metrics to incorporate. Our team started by following this recommendation, but over time, we have moved away from collecting errors for two reasons.

First, the inter-rater reliability of designating errors is not high. It is difficult to assess errors consistently across different researchers, especially when reviewing substantially different platforms, products, and scenarios. This lack of consistency makes it more difficult to realize the goal of the single metric – consistent comparison over time or across competitors.

Second, there is ambiguity in how an error rate should be calculated. Typically, the number of errors is to be divided by the number of possible errors. The number of possible errors is tricky (conceivably it could be almost infinite), and the recommendation [5] is to take the number of errors +1 as the denominator.

This is obviously a choice of convenience, acknowledging that it is difficult to know how many errors could possibly be. This reduces the reliability of this error rate calculation even further, yielding a metric with compounding sources of noise. As a result of these limitations, we stopped using error rates as a core usability metric. Other SUM users similarly avoid errors [10], and Jeff Sauro has acknowledged the difficulty of reliably collecting errors and suggested a three-component SUM. [8]

3.2 What is a good time on task?

Time on task is a problematic metric as it is very difficult to objectively establish what is a 'good' or 'usable' amount of time for a task across users. A common practice is to use a reference time multiplied by some constant to give an acceptable range of times. The reference time is typically either an expert's time on task, or sometimes the fastest participant. Multipliers could be anything... 1.5, 2, and 3 have all been suggested and tried. The unfortunate reality is that these multipliers are arbitrary and do not objectively delineate usable or unusable time-on-task.

A mental exercise can illustrate its limitations. Let's say we use expert time x 3 for our 'spec time.' If our task is very simple, say "navigate home" on a website with a clear title link, this task is a single click and could be completed in a second or less by an 'expert.' That would make our usable threshold 3 seconds. If a participant were to happen to sneeze, or reread instructions, or sip a drink, they could take well over 3 seconds without any error or difficulty experienced. The participant could be quite positive but end with an "unusable" score. On the other hand, what if our task is very long, such as completing a first run experience on a new platform? Perhaps it takes an 'expert' 15 minutes to complete the task... 45 minutes would not be an acceptable time for what was intended to be a 15-minute experience, no matter how novice the user. The 3 * expert time spec limit does not satisfy a wide range of possible tasks.

If an arbitrary multiplier cannot support a wide range of tasks, perhaps an algorithmic one could. Unfortunately, finding a spec limit using standard deviation and/or z-score has been difficult. Using a particular z-score as a cut-off effectively creates a pre-defined failure rate. For example, saying any time-on-task with a z-score of -2 or less effectively says that you will always have a 3% failure rate. "Standardized" failure rates are not very helpful in comparing two experiences.

Alternatively, trying to base a threshold on standard deviation or variance will penalize tasks with multiple success paths (resulting in high satisfaction and success, but spread out time-on-task) and also prove less rigorous for short tasks that naturally cluster closer together, like the aforementioned 'navigate home' example.

Overall, our dataset has shown time to be a fickle metric indeed. The simple correlation of the average time on task (transformed using natural logarithms to a more normal distribution) and task completion rate is only 0.23 for the 500-task data set, and 0.30 for satisfaction. (This result contrary to Sauro & Kindlund [1] and Lewis [2]) Tellingly, the correlation of time on task and perceived time (as measured with a Likert scale) is only 0.35, showing that the length of time people experience did not strongly tie to how long they felt it took. Intuitively, we understand scenarios where the time does not correlate well: some participants give up early and move on from a difficult task, while others struggle for a while but eventually succeed. Novice users often must deal with discoverability and learnability, or have varied knowledge of shortcuts, increasing variance of time somewhat independent of completion rate. Though efficiency is a stated goal within usability, it is not clear from this and other results [9] how strong of a contributor it is to the perceived usability of a product.

As a result of these limitations, we have moved away from using time-on-task as a component of a single usability metric, instead preferring to only compare it directly with similar experiences in pairwise tests.

3.3 Averaging of Satisfaction

Another issue encountered is the averaging of subjective metrics into one. Originally, Sauro and Kindlund recommended averaging three subjective scores – perceived time, perceived ease, and satisfaction. When averaging subjective ratings, great care must be taken to ensure the rating scales are on similar basis. For example, there may be two ways to ask a participant how they perceived time on a 5-point Likert scale:

- 1 Very slow… 3 Neither fast nor slow… 5 Very Fast
- 1 Much slower than expected… 3 As expected… 5 Much faster than expected.

In the first, "3 Neither fast nor slow" is difficult to judge as "usable" or not, and so a spec limit of 4 may be used. For the second, a 3 rating translates to "As expected." Meeting expectations is likely "usable," and being "faster than expected" is a harder goal to attain than being "fast" in many cases. For example, in the "navigate home" example, users might rate it "5 Very fast" for the first scale, but "3 As expected" for the second. These two ways of anchoring a question result in different spec limits for being 'usable.'

If subjective metrics are averaged, but have different implied spec limits (e.g. '3 As expected' for time, but '4 Somewhat satisfied' for satisfaction) the average of the metrics will need the average of the spec limits to be accurate – this is often not an intuitive number to report.

Table 1: The selected thresholds (out of 5) for subjective metrics based on their labels.

Metric	Threshold	Value Label
Perceived Time	3	'As Expected'
Perceived Difficulty	4	'Somewhat Easy'
Satisfaction	4	'Somewhat Satisfied'

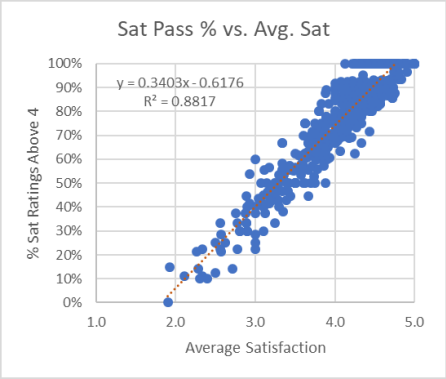

Figure 1 A scatter plot of tasks showing average satisfaction against the % of high ratings.

3.4 Communication Challenges

The above issues combine to render the SUM into a highly sensitive and difficult to explain variable. For example, a task could have very high success, very high satisfaction, but the time-on-task fall outside of the 3x multiplier window (as we noted above, this is common on very short tasks). If we discard errors and track only three metrics, this may result in a SUM as low as 66%, despite having perfect success and satisfaction. Alternatively, if subjective measurements are used with different spec limits, averaging them all together could result in false negatives, lowering the SUM again. These scenarios can create difficulties in communicating with stakeholders – it may appear that a task has very high completion rates, or even high satisfaction scores, but the summative metric shows a value that implies a large need for improvement.

These communication issues are further complicated by the statistical processes used to generate the SUM, specifically the normalization by z-score steps, and further high-level reasoning to compute the necessary confidence intervals. [10] Though perhaps not overly difficult concepts, any level of statistical processing often adds a layer of obfuscation to a metric for stakeholders untrained in statistics. The net result is a metric that may yield unexpected results and trying to unpack the causes of a lower score can involve significant statistical reasoning. Frequently, there is not enough time when presenting research results to unpack the nuanced causes of score changes, which reduces its usefulness in comparing scores.

Over the course of our benchmarking program many of the above issues were addressed in one form or another, but at the end of a significant period of testing, a simpler metric was deemed necessary to facilitate the stakeholder communication needed to have the metrics be impactful.

3.5 A Simpler Metric

It is useful to review our goals for a single usability metric: to be a valid, reliable metric that is simple to calculate (and therefore explain to stakeholders) that facilitates the comparison over time or across products at a high level. Tullis and Albert [6] suggest several such metrics in Measuring the User Experience, and for our initial pivot, we turned to the simplest and most straightforward of them all – combining metrics based on target goals.

Like SUM, we have several component metrics, and each is compared to a threshold for "good usability." Each metric becomes a binary variable – did it pass the threshold or not? Each binary variable can be unioned together to yield a single binary value for a given user on a given task. All users can be averaged for a task to give a usability rate for the task.

For this metric, we can reuse the metrics collected throughout the history of this project: task success, perceived time, perceived ease, and satisfaction. Errors are not collected due to reliability issues, and time-on-task is not incorporated due to the difficulty in establishing an objective threshold for a good time-on-task. In this particular case, each variable is given equal weighting, but different weighting schemes can be adopted as desired.

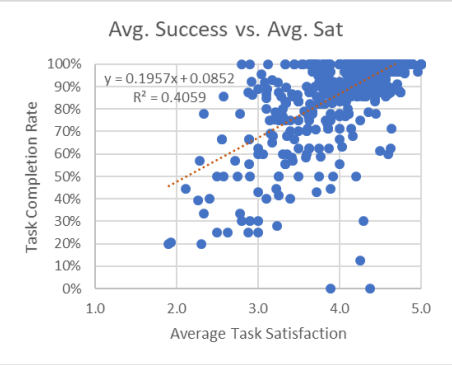

Figure 2 A scatter plot of tasks, illustrating correlation between satisfaction and success.

Table 2 Target passing rates for the single metric, as correlated with passing thresholds for the component metrics.

Metric (Threshold)	Avg Single Metric	Adj Wald Single Metric
Completion (87%)	61%	58%
Perceived Time (3)	54%	54%
Perceived Difficulty (4)	59%	57%
Satisfaction (4)	60%	58%

Like SUM, this metric requires thresholds to establish a good score. Likert scale rating thresholds are developed from the anchoring text of the choices as shown in Table 1. Combined with whether the participant completed the task, we have metrics to cover the efficiency, efficacy, and satisfaction aspects of usability [7]. For a given user on a given task, the single metric is a '1' if they complete the task and rate each subjective metric above its threshold. The average across all users shows the percentage of users who found the task 'usable' – a measurement more easily explained with stakeholders. Internally, we call it a usability score.

Our dataset lets us go a step further and establish what metric ratings can be considered 'good' for usability. Let us start with an assumption, that we want users to be at least somewhat satisfied on average. This starts us with a target, to have an average satisfaction rating of '4'. Due to the restricted nature of a 5-point Likert scale, this is a fairly high requirement: a task average satisfaction of 4.0 correlates with 74% of participants rating a 4 or 5, on average (Figure 1.)

Correlations among the subjective metrics are very high, and so the other two have very similar results: a perceived difficulty of 4 corresponds to a 74% pass rate, and perceived time (where a rating of 3 is 'as expected') corresponds to 71% rating the metric as 3 or above.

With these subjective metric targets, we can objectively determine a target completion rate to be judged as 'usable.' The subjective metrics are fairly well correlated with task completion (though not overwhelmingly so.) For the dataset in question, we find task completion to have correlations of **0.56, 0.66, and 0.64** with perceived time, perceived difficulty, and satisfaction (distance correlations of 0.55, 0.64, and 0.62.) These correlations are roughly in-line with those reviewed by Lewis [2] and Hornbæk [3]. Though not perfect, we can develop a regression for each metric to determine a completion rate that correlates with our positive subjective rating labels.

For a simple average completion rate, our three subjective metric thresholds correlate to a completion rate of **87%, 87%, and 86%** (Figure 2.) If we calculate the completion rate using the Adjusted Wald interval mean instead, our target for completion falls to **80/80/79%.** This shows a high bar of success is necessary to yield high satisfaction products, higher than sometimes suggested.

We can repeat the same process with the calculated single metrics to find a comparable level of "overall usability." The resulting thresholds shown in Table 2 are rather low, showing the challenge in getting all metrics to align together for any given participant, as well as the information lost in converting to a binary metric. Due to the interplay of the component metrics, a score as high as 75% is needed to ensure all component metrics reliably pass their thresholds.

Higher standards can be employed by teams as desired, and individual metric thresholds provide the basis for discussion. These metrics are easily calculated and explained, and when combined with confidence intervals to illustrate the reliability of surpassing the thresholds, provide usability data that can be quickly and effectively communicated with product stakeholders.

4 CONCLUSION

A single usability metric has value in clearly communicating an overall level of usability to stakeholders and facilitating the comparison over time, over products, or to defined objectives. Our experience with SUM found that the process needed simplification in order to effectively communicate with our stakeholders and so have impact on product. By using already well understood methods and generalist-level techniques, we have been able to utilize an effective single usability metric and develop parameters for usability reporting.

REFERENCES

[1] Sauro, J., & Kindlund, E. (2005). How long should a task take? identifying specification limits for task times in usability tests. In *Proceeding of the Human Computer Interaction International Conference (HCII 2005)*, Las Vegas, USA.

[2] Sauro, J., & Lewis, J. R. (2009, April). Correlations among prototypical usability metrics: evidence for the construct of usability. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1609-1618). ACM.

[3] Hornbæk, K., & Law, E. L. C. (2007, April). Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 617-626). ACM.

[4] Sauro, J., & Kindlund, E. (2005, April). A method to standardize usability metrics into a single score. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 401-409). ACM.

[5] Sauro, J., & Kindlund, E. (2005, July). Using a single usability metric (SUM) to compare the usability of competing products. In *Proceedings of the Human Computer Interaction International Conference (HCII)*.

[6] Albert, W., & Tullis, T. (2013). Measuring the user experience: collecting, analyzing, and presenting usability metrics. Newnes.

[7] ANSI (2001). Common industry format for usability test reports (ANSI-NCITS 354-2001). Washington, DC: American National Standards Institute.

[8] Jeff Sauro (2012). 10 things to know about the single usability metric (SUM). Retrieved December 10, 2018 from https://measuringu.com/sum/

[9] Lallemand, C. & Koenig, V. (2017). Lab testing beyond usability: challenges and recommendations for assessing user experiences. *Journal of Usability Studies*, 12(3), 133-154

[10] Bradner, E., & Dawe, M. (2008). Parts of the sum: a case study of usability benchmarking using the sum metric. In UPA international conference (S64)