Is the 'Figure of Merit' Really that Meritorious?

Jarinee Chattratichart[1] and Gitte Lindgaard[2]

[1]Faculty of Computing, Information Systems and Mathematics, Kingston University, London, United Kingdom
[2]HOTLab, Psychology Department, Carleton University, Ottawa, Canada

Abstract. Studies comparing performance of Usability Evaluation Methods (UEMs) led to three standard metrics, namely, validity, thoroughness, and effectiveness, calculated from lab-based usability test results. The effectiveness metric, E = T x V, was proposed as the 'figure of merit' [7] that would give a balanced account of validity and thoroughness. This paper provides an analysis of the formula to caution future researchers and usability practitioners against its use, proposes an alternative formula, and discusses the limitations of the common baseline approach to UEM comparison.

Keywords: comparative usability evaluation, UEM, metrics

1 Background

In the early 1990s when usability inspection methods (UIMs) were introduced as quicker and cheaper alternatives to usability testing, there was a surge of comparative studies of such methods. However, these studies yielded inconclusive results due to poor research design [6], a lack of standard measures, and a baseline for fair comparison [13]. Subsequently, three UEM performance metrics based on a common baseline were introduced and employed in Sears' [13] study. These metrics, namely, validity, thoroughness, and reliability, are computed from results of a usability test of the same interface using the following formulas:

$$\text{Thoroughness, } T = \frac{\#\text{ of Real Problems Found}}{\#\text{ of Real Problems that Exist}}$$

$$\text{Validity, } V = \frac{\#\text{ of Real Problems Found}}{\#\text{ of Issues Identified as Problems}}$$

$$\text{Reliability, } R = \max(0, R_{temp}), \text{ where } R_{temp} = \frac{1\text{-stdev(\# Real Problems Found)}}{\text{average (\#Real Problems Found)}}$$

Hartson, et al. [7] support Sears' common baseline approach to UEM comparison but raise a concern that neither T nor V alone "is sufficient for *[assessing]* UEM effectiveness" (p.394). They propose a 'figure of merit' for measuring UEM effectiveness that takes into account both T and V, "reflecting a more balanced overview of UEM performance" (p.394). The 'figure of merit' is defined as Effectiveness, E = TxV. Additionally, by way of analogy to the concept of precision (P) and recall (R) in Information Retrieval (IR) and Natural Language Processing

(NLP), they propose "a weighted measure for UEM performance" (*F*) using the *F*-measure formula [10] that is a well-known effectiveness metric used in those fields, but replacing the P and R variables in the formula by V and T, respectively:

$$F = \frac{1}{\alpha\,(1/V)+(1-\alpha)(1/T)}$$

where, α is "the factor used to set the weighting between thoroughness and validity".

After a decade of debate and uncertainty of how to compare UEMs fairly, that work [7, 13] is very valuable. Many studies have employed T, V, and/or E [1, 2, 3, 4, 7, 9, 13]. We provide arguments for and against the E formula, propose a new metric for E, and discuss the use of performance metrics to compare UEMs.

2 Arguments for and against the E metric

The denominator of the T formula above has a constant value for a usability test. T is thus high when the number of predicted real problems is high. Yet, to achieve the latter, more evaluators must inspect the interface [12]. Increasing the number of evaluators allows more false alarms and in turn reduces V [4]. However, to increase V, requires fewer evaluators. More problems are therefore missed leading to a lower T. In short, T and V affect each other negatively. It is therefore unjustified to assess or compare UEMs using T or V alone. Hartson et al.'s [7] reason for a 'figure of merit' that gives an overall effect of T and V is thus justified. However, they did not explain why E should be equal to T times V, nor did they show the derivation of this formula.

The formula has since been used unquestioned. Empirical evidence from [1, 2, 3, 4, 9] shows that the E value is *'capped'* (i.e. it is always lower than the lower value between T and V). This raises our concern because if E is intended to reflect "a balanced overview of UEM performance" by taking into account both T and V [7], why should the balanced value be lower than the lower value of the two? Could these results be mere coincidence?

The above results were not coincidental. Mathematically, the E value is expected to be *'capped'*. T and V values are ratios, ranging from 0 to 1 [13]. When $0 < V < T$, and if we multiply both sides of the equation, $T \leq 1$, by V we get $TxV \leq V$. Replacing TxV by E results in $E \leq V$. Similarly, when $0<T<V$, it can be shown that $E \leq T$ by starting the same process with $V \leq 1$. There is no question that the E value is *'capped'*, but why should it be, especially if it aims to reflect a balance between T and V? We could not find a direct answer in [7], nor in the literature. However in [7], the *F*-measure was proposed as "a weighted measure for UEM performance" for the same purpose as E, in which case the two formulas should yield somewhat similar values. Our next question is whether the *F* value is *'capped'* also. A quick demonstration shows that this is not the case. Assuming that T and V have equal weight, $\alpha = 0.5$. Replacing this in the *F* formula yields $F = 2TV/(T+V)$, in other words, the harmonic mean (HM) of T and V. A mean value of A and B will always fall between A and B. Hence, the *F* formula gives a UEM effectiveness value that falls between T and V. Why, then, should the UEM's overall performance value be both between T and V (the *F* formula) and capped by the lower value between T and V? Which one should

usability practitioners and researchers use if they wish to use a single composite metric to assess or compare UEMs?

The last argument against the E and in favor of the *F* formula is that E is ill-behaved while *F* is well-behaved. A standard metric should be well-behaved to avoid violation of statistical assumptions commonly required for data analyses. This is especially relevant to comparative UEM studies. A well-behaved variable has a normal distribution and no outliers [5]. We ran a simulation of 2501 values of E and HM of T and V (or *F* when $\alpha = 0.5$) from the 2501 (T, V) pairs of all possible combinations of T and of V values ranging from 0.02 to 1 with equal increments of 0.02 from one pair to the next. This yielded a total of 2500 pairs to which the (0, 0) pair was added. The (0, n) and (n, 0) pairs (where $0 < n <= 1$) were excluded because if one of the two metrics (T, V) is 0, i.e. no real problem exists or no real problem is predicted; the value of the other metric must therefore also be 0. The results revealed a positively skewed distribution and outliers for E, but a normal distribution, with no outliers for *F*. Skewness, mean and median were 1.0, 0.2 and 0.2 for E and 0.3, 0.4 and 0.4 for *F*, respectively. Hence, E is ill-behaved and in one half of the cases the E values do not exceed 0.2; on average, the value was only 0.2 on a scale 0 to 1.

3 Discussion

The above analysis suggests that the HM or weighted HM of T and V is safer than the E formula for giving an overview and a balanced value between T and V. However, the denominator of HM of T and V, T+V, violates a rule for addition. In mathematics, addition and subtraction can only be performed on like terms or same unit of measurements. Although both T and V are proportions, they differ semantically and their denominators are derived from different units. For T, the denominator is all real problems that exist; for V, it is all predicted problems including both real problems and false alarms. If the 'figure of merit' is to have a value between T and V, their geometric mean (GM) is a better option than HM because GM of T and V = $\sqrt{(TxV)}$ and multiplying unlike terms in mathematics is allowed. The above GM simulation revealed a normal distribution with no outliers and a mean and median of 0.45 and 0.45, respectively. Our proposal for a new 'figure of merit' is hence, GM or weighted GM: $E = \sqrt{(TxV)}$ or $E = T^{\alpha} \times V^{(1-\alpha)}$, where α is the weighted ratio of T.

Using a common baseline approach to compare UEMs should only be done within the same study, using the data from the same usability tests. This is because it is unlikely that a usability test will reveal all problems that exist or that different tests would yield the same results [11], making comparisons across studies unfair.

How does this analysis affect previous research using this approach? It does not affect conclusions about UEM performance as all figures are relative and have a common baseline. Yet, with E values calculated using the new formula would be higher than those published, fall between T and V, and they do not violate statistical assumptions commonly required for data analysis.

Another limitation of this common baseline approach is that it is performance-focused. However, there are many aspects of usability to measure and, at present, the choice, validity and reliability of usability measures used in usability tests is a

pressing issue awaiting future research [8]. Performance metrics alone are not sufficient for assessing UEMs. Future studies should also compare UEMs on other usability aspects such as retention, learning, user satisfaction and perception.

4 Conclusion

We have presented arguments against future use of the E formula and suggested that the geometric mean of T and V be used instead. Limitations of the common baseline approach to UEM comparison and future directions were also discussed.

References

1. Andre, T. R., Hartson, R., & Williges, R. C. (2003). Determining the effectiveness of the Usability Problem Inspector: A theory-based model and tool for finding usability problems, *Human Factors*, *45*(3), 455-482.
2. Chattratichart, J., & Brodie, J. (2004). Applying user testing data to UEM performance metrics. In *CHI '04 extended abstracts on Human factors in computing systems* (pp 1119-1122). ACM.
3. Chattratichart, J. & Lindgaard, G. (2008). A comparative evaluation of heuristic-based usability inspection methods. *In CHI '08 extended abstracts on Human factors in computing systems* (pp 2213-2220). ACM.
4. Cockton, G. & Woolwich, A. (2002). Sale must end: Should discount methods be cleared off HCI's shelves? *Interactions*, *9*(5), pp 13-18. ACM.
5. Gray, A. R., & MacDonell, S. G. (1997). A comparison of techniques for developing predictive models of software metrics, *Information and Software Technology*, *39*, 425-437.
6. Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods, *Human-Computer Interaction*, *13*, 203-262.
7. Hartson, H. R., Andre, T. S., & Williges, R. W.(2001). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction*, *13*, 373-410.
8. Hornbaek, K. (2006). Current practice in measuring usability: Challenges to usability studies and research, *International Journal of Human-Computer Studies, 64*, 79-102.
9. Howarth, J., Smith-Jackson, T., & Hartson, R. (In Press). Supporting novice usability practitioners with usability engineering tools, *International Journal of Human-Computer Studies*, Available online 27 February, 2009.
10. Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.
11. Molich, R., Ede, M. R., Kaasgaard. K., & Karyukin, B. (2004). Comparative usability evaluation, *Behaviour & Information Technology*, *23*(1), 65-74.
12. Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of INTERCHI Conferences on Human Factors in Computing Systems* (pp. 206-213). Amsterdam: Association for Computing Machinery.
13. Sears, A. (1997). Heuristic Walkthroughs: Finding the problems without the noise. *International Journal of Human-Computer Interaction, 9*(3), 213-234.