RELIABILITY OF SEVERITY ESTIMATES FOR USABILITY PROBLEMS FOUND BY HEURISTIC EVALUATION

Jakob Nielsen
Bellcore
445 South Street
Morristown, NJ 07962-1910
nielsen@bellcore.com

ABSTRACT

Ratings from single evaluators are very unreliable when usability specialists judge the severity of usability problems found by heuristic evaluation, but the mean severity rating from four evaluators gets within half a rating point of the true severity 95% of the time. Also, the evaluators do agree that usability problems found by heuristic evaluation are all real problems even though each rater had originally only identified a small proportion of the problems.

Keywords: Severity, Prioritizing, Evaluation Methods, Usability Problems, Judgments.

INTRODUCTION

The data presented here was collected as part of a heuristic evaluation [1][2] of a prototype user interface for a software product for very specialized telephone company staff. The interface was subjected to heuristic evaluation by eleven usability specialists who found forty usability problems in the interface.

SEVERITY JUDGMENTS

After the heuristic evaluation, the evaluators were given a questionnaire listing all forty usability problems (including the problems found by other evaluators than themselves) and were asked to rate their severity on a 0–4 scale, with 0 indicating that the "problem" was not a usability problem at all, 1 that it was a cosmetic problem, 2 that it was a minor problem, 3 that it was a major problem, and 4 that it was a usability catastrophe. As a first major result, all forty problems were rated higher than zero by a majority of the evaluators, indicating that they were indeed usability problems, even though the average evaluator had only found 29% of the problems him/herself in the original evaluation sessions. For three of the forty problems, however, as many as 27–36% of the evaluators disagreed that they were usability problems.

In addition to the heuristic evaluation, a small user test was conducted with four representative test users. The correlation between the number of users observed having a problem and the heuristic evaluators' mean severity rating of that problem is .46 which is reasonably high (and significant at $p<.01$), lending some credence to the validity of the severity ratings. Of course, there are other factors to take into account when calculating the severity of a usability problem than just how many users experience the problem, but the nature of the prototype prevented the collection of traditional task performance measures.

RELIABILITY OF SEVERITY JUDGMENTS

The average correlation between any two evaluators' severity ratings of the same problems is .24. Kendall's coefficient of concordance between the eleven evaluators is W=.31, which is statistically significant (χ^2=132.3, df=39, $p<.01$) and thus indicates that the agreement is not just chance. Also, of the 55 pairwise comparisons between evaluators, only four have negative correlations, whereas the remaining 51 are positive.

Even though the statistics indicate better than random agreement between evaluators, the inter-rater reliability is still very low compared to the standards of most respected rating methods. Basically, the reliability of the severity ratings from single evaluators is so low that it would be advisable not to base any major investments of development time and effort on such single ratings. On the other hand, the better-than-random agreement between evaluators means that it is possible to use the mean of the severity judgements from several evaluators and get much more reliable results. It is a fairly simple task for an evaluator to produce severity ratings for an interface which is known to that evaluator from a heuristic evaluation session or otherwise, and the evaluators in the case study presented here spent about half an hour each on doing so. Therefore, it would seem reasonable to ask for severity judgements from all or at least most evaluators.

The Spearman-Brown formula for estimating the reliability of combined judgments from several evaluators is

$$r_{n\text{-}n} = \frac{n \cdot r_{1\text{-}1}}{1 + (n-1) \cdot r_{1\text{-}1}} \qquad \text{(EQ 1)}$$

and Figure 1 shows a plot of the way the reliability of the mean severity estimate increases as more evaluators are added.

The standard error of measurement for the true underlying value of a rating derived as the mean of n ratings is

$$\sigma_\infty = \sigma_n \cdot \sqrt{1 - r_{n\text{-}n}} \qquad \text{(EQ 2)}$$

where $r_{n\text{-}n}$ is the reliability of a group of n raters and σ_n is the standard deviation of the mean of the n ratings which again is

$$\sigma_n = \sigma_1 / \sqrt{n} \qquad \text{(EQ 3)}$$

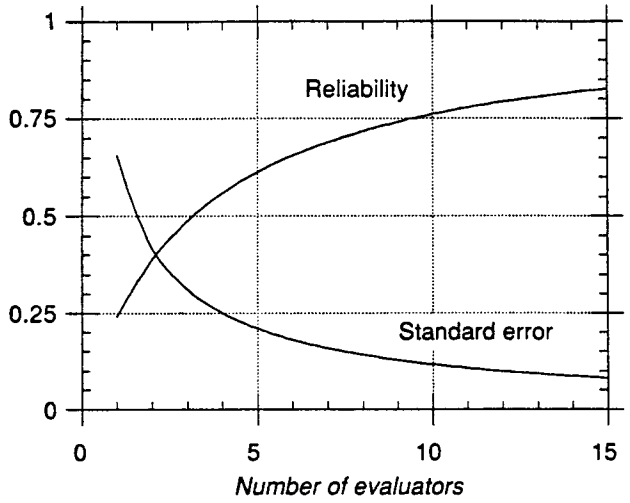

Figure 1 The upper curve shows the reliability of the severity estimates as a function of the number of evaluators used. The lower curve shows the standard error of measurement (in rating units from the 0–4 scale) for the underlying true mean severity value when measured by the mean of the ratings from that number of evaluators. Values for more than one evaluator are plotted according to the Spearman-Brown formula.

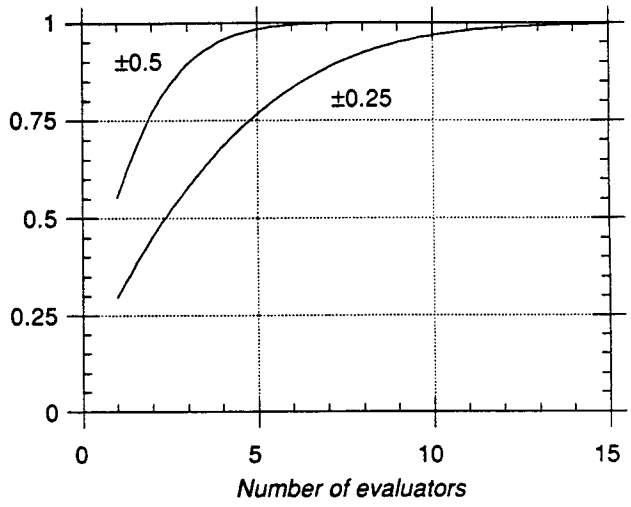

Figure 2 Probability of having the mean severity rating from sets of one through fifteen evaluators be within ±0.5 and ±0.25 rating units of the true severity of a usability problem.

if the individual ratings can be assumed to be independent variables. This assumption of independence obviously only holds if the evaluators perform their evaluations separately and do not discuss the usability problems before giving their severity judgments.

Combining equations 1 to 3 gives the standard error of measurement shown in Figure 1. Since the severity judgments are actually fairly close to following a normal distribution, we can use the normal distribution to calculate confidence intervals for the severity estimates. The standard deviation of the complete set of severity ratings adjusted for means is 0.75 which can be used for a general estimate of σ_1.

Because of the low reliability of the severity ratings and high standard deviation for the individual ratings, the probability of having a single evaluator provide an estimate that is within ±0.5 rating units of the true severity of a problem is only 55%. In other words, almost half of the time, the absolute rating will be substantially different from the "true" rating (that would result from having an infinite number of evaluators). On the other hand, combining the estimates of several evaluators considerably improves the confidence intervals for the mean estimate. With just two evaluators, one has a 77% chance of getting within ±0.5 of the true severity, with three evaluators, the chance goes up to 91%, and with four evaluators, the chance of getting within ±0.5 is 95%. Figure 2 shows how the probability of getting within ±0.5 and ±0.25 goes up with the number of evaluators.

Severity estimates for usability problems have two possible practical applications: First, one needs to know the absolute need for usability improvements in order to determine the need for continued usability efforts and consider the extent to which an increased usability engineering budget or delays in product introduction is warranted. Second, given a specific budget for usability activities and the need to move on in the product lifecycle within a given timeframe, one needs to set relative priorities for which problems to fix first.

For estimating the absolute severity of the usability problems in an interface, it may be sufficient to estimate the severity of each usability problem with an uncertainty of ±0.5 rating units. The length of a ±0.5 interval is of course one unit, and it would be difficult to interpret the meaning of subjective usability severities with a much finer resolution than that. For example, any cost-benefit estimate of the potential fixing of usability problems has to include estimates of the programming effort needed to implement a redesign as well as an estimate of the usability of the new design. Neither value can be estimated with great precision anyway, so it would be a wasted effort to measure the severity of the usability deficiencies in old design with extremely narrow confidence intervals. It would be valuable also to have externally valid data to calibrate the subjective rating scale with respect to the economic impact of fixing usability problems of various severities.

CONCLUSION

Based on these considerations, one can definitely conclude that severity ratings from a single evaluator are too unreliable to be trusted. As more evaluators are asked to judge the severity of usability problems, the quality of the mean severity rating increases rapidly, and ratings from three or four evaluators would seem to be satisfactory for most practical purposes.

Acknowledgment

The author would like to thank Tom Landauer for valuable help with the statistical analysis of the data presented here.

References

1. Nielsen, J. Finding usability problems through heuristic evaluation. *Proc. ACM CHI'92* (Monterey, CA, 3–7 May 1992).
2. Nielsen, J., and Molich, R. Heuristic evaluation of user interfaces. *Proc. ACM CHI'90* (Seattle, WA, 1–5 April 1990), 249–256.