COMPARING STUDIES THAT COMPARE USABILITY ASSESSMENT METHODS: AN UNSUCCESSFUL SEARCH FOR STABLE CRITERIA

Michael J. Muller,[1] Tom Dayton,[2] and Robert Root[3]

Bellcore
Piscataway NJ 08854 US

ABSTRACT

Four studies that compared inspection methods with usability testing were re-analyzed using six distinct criteria for the superiority of one method to another. Each study's own results were found — to a greater or lesser extent — to be in *internal* conflict when examined across the six criteria. These analyses, added to the well-known contradictions *across* the studies, argue that any conclusions regarding overall superiority of one method with respect to another are premature. They also lead to questions regarding the selection of criteria.

KEYWORDS: Usability testing, inspection methods, comparisons of methods, user centered design.

In 1990, Jeffries, Miller, Wharton, and Uyeda (1990) sparked a controversy regarding the relative efficacy of one usability assessment method to another. Using benefit-cost analyses, they showed a 12-to-one superiority for an expert inspection method (heuristic evaluation) over usability testing. Jeffries and colleagues were careful to state that good usability practice would involve *both* inspection methods and usability testing. Unfortunately, their results have often been misinterpreted as arguing against usability testing — to such an extent that Jeffries and Desurvire (1992) felt it necessary to print a clarification of the on-going need for usability practices that *combine* elements of inspection with usability testing.

Several other studies examined the same underlying question, albeit from somewhat different perspectives. Karat, Campbell, and Fiegel (1992) and Desurvire, Kondziela, and Atwood (1992) found differences between inspection methods and usability testing. However, the patterns of differences that they found were in contradiction with those of Jeffries and colleagues. Unpublished Bellcore data of Nielsen, Norwood, and Muller also provided a more limited set of comparisons.

Because each research group was scrupulous in publishing detailed data, we were able to re-analyze the studies. We considered five criteria by which to compare usability assessment methods:

- **Raw yield.** How many *unique classes of usability problems* were discovered by each method?
- **Raw yield weighted by opportunity.** What was the raw yield *per participant hour* (i.e., per hour of opportunity to discover problems) for each method? This criterion gives a more reasonable estimate than raw yield, because it avoids distortions that may be due to differing numbers of user testing hours vs. expert inspection hours, etc.
- **Refined yield.** What proportion of *severe* problems were found by each method? This criterion may be more useful to developers than simple lists of problems, because developers can't fix every problem, and they need a method that has a higher detection rate for serious problems.
- **Benefit-cost.** What was the average cost (in terms of total human hours involved) in finding each problem, for each method? This criterion is attractive from an engineering and/or management perspective, of allocating scarce resources where they will do the most good.
- **Uniqueness.** Which methods are more likely to find problems that are undiscovered by other methods?

In addition, we noted that there appeared to be different assumptions about the ratio of participant hours to analyst hours, as practiced in the Jeffries and Karat studies. Jeffries and colleagues assumed that inspection methods required virtually no post-session analysis time, whereas Karat and colleagues assumed that inspection methods required considerable post-session analysis time. We therefore made a first approximation *what-if* analysis, to attempt to see what each study's results would have looked like if they had been analyzed using *the other study's* assumptions.

Our analyses are summarized in Figure 1. In a paper of this brevity, we cannot discuss them in detail. We note simply that Figure 1 shows clearly two types of inconsistencies:

- **External inconsistencies.** As is known, the studies' outcomes are in disagreement with one another.

1. Current address: U S WEST Advanced Technologies, 4001 Discovery Drive, Boulder CO 80303 US +1-303-541-6564 michael@advtech.uswest.com
2. Bellcore — Room RRC-1H226, 444 Hoes Lane, Piscataway NJ 08854 US +1-908-699-6843 tdayton@ctt.bellcore.com.
3. Bellcore — Room RRC-1H226, 444 Hoes Lane, Piscataway NJ 08854 US +1-908-699-7763 broot@ctt.bellcore.com.

- **Internal inconsistencies.** Within each study, the different criteria often lead to radically different orderings among the methods. Similarly, the *what-if* analyses show that simple differences in staff-hour assumptions may *reverse* the relative efficacy of the methods.

These observed inconsistencies lead us to one conclusion and one question (or one set of questions). The conclusion: There is no stable pattern of superiority of one method over another, *even within a single study*. Any claim that one method is more effective than another depends more upon the criterion of effectiveness, than upon any invariant pattern within the data. The question(s): We urge practitioners and researchers to consider conjointly both the validity and the social consequences of the criteria they use to compare methods. Is benefit-cost more important than uniqueness? Is user participation in product assessment more important (in long-term business consequences, as well as moral and political issues) than economic issues that become invisible as soon as the study is over? We doubt that these questions lead to a single set of answers. We believe that a discussion of our diversity in answers would prove valuable.

ACKNOWLEDGEMENTS

We thank Jakob Nielsen, Gay Norwood, and Deondra Robinson for contributing to the reanalysis of the Bellcore data.

REFERENCES

[1] Desurvire, H., Kondziela, J., and Atwood, M. (1992). What is Gained and What is Lost when using Evaluation Methods other than Usability Testing? Paper presented at Human Computer Interaction Consortium.

[2] Jeffries, R., and Desurvire, H. (1992). Usability Testing vs. Heuristic Evaluation: Was There a Contest? *SIGCHI Bulletin* 24(4), 39-41.

[3] Jeffries, R., Miller, J., Wharton, C. and Uyeda, K. (1991). User interface evaluation in the real world: A comparison of four techniques. In *Proceedings of CHI'91*. ACM: New Orleans LA, 119-124.

[4] Karat, C. M., Campbell, R., and Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of CHI '92*. Monterey CA: ACM, 397-404.

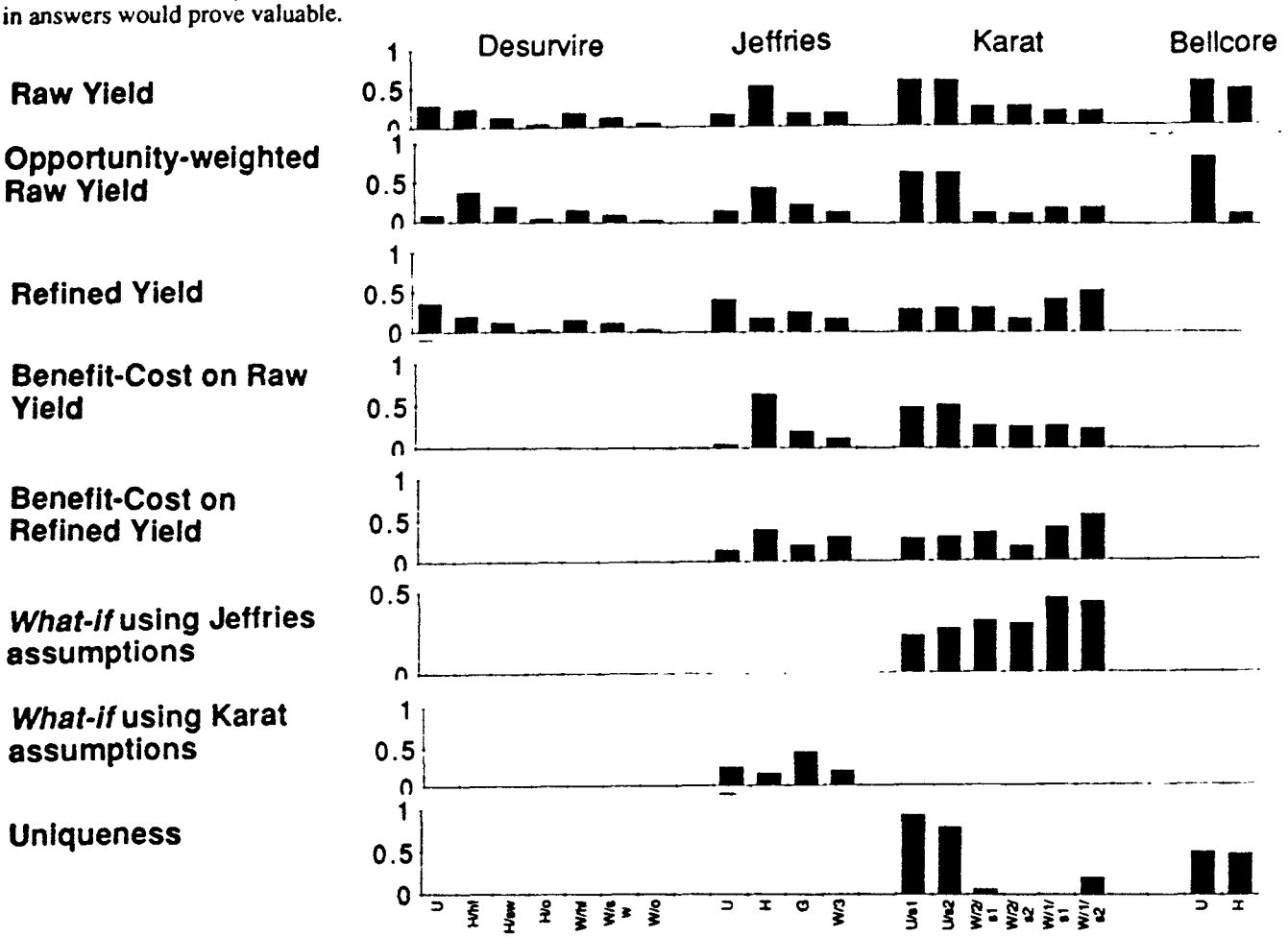

Figure 1. Comparisons of the outcomes using five criteria on the four studies. U=usability testing, H=heuristic evaluation, W=cognitive walkthrough, G=guidelines; /hf=human factors participants; /sw=software engineering participants; /o=other participants; /1,/2,/3=size of group conducting the inspection, /s1=test system 1, /s2=test system 2. Some studies did not provide data for calculating certain criteria.