© Studio Archeology/Comstock, Inc.

Sale Must End: Should Discount Methods be Cleared off HCI's Shelves?

If you ask someone outside the Human-Computer Interaction (HCI) field about usability, many will mention the "classic" discount methods popularized by Jakob Nielsen and others. Discount methods have the appeal of seeming easy to do, and, more importantly for business, being inexpensive. This is especially attractive to smaller startup companies with low budgets. But are discount methods really too risky to justify the "low" cost? This month's business column authors think so, based on their research and experience. Indeed, they believe that these discount methods may actually backfire and end up discrediting the field. Following a lively discussion on the CHI-WEB listserv, we asked them to explain what they see the risks to be, and what they believe we, as a profession, can and should do about it.

— David Siegel and Susan Dray

Gilbert Cockton and Alan Woolrych
School of Computing and Technology, University of Sunderland, UK
Gilbert.Cockton@sunderland.ac.uk

From Cost Cutting to Cost-Benefit

The value of so-called "discount" usability methods is much discussed. These methods cut corners in the hope that "some usability is better than no usability." They cut costs by reducing demands on the critical resources of time, facilities, cash, and skill. For example, discount user testing saves time and cash by testing up to five users, and reduces demands on skill by allowing limited test planning, simple test execution and "lite" analysis of usability "results." Inspection methods discount by examining only some aspects of usability problems, thus saving on analysis time and analyst skill requirements (Cockton, Lavery and Woolrych 2002).

Definitions

Predicted problem set = the merge of all analyst predictions
Actual problem set = the merge of all empirically derived problems

Hit = successful prediction (in predicted and actual problem sets)
Miss = unpredicted problem (in actual problem set only)
False Alarm = unsuccessful prediction (in predicted problem set only)

$$\text{Thoroughness} = \frac{\text{hits}}{\text{hits+misses}}$$

$$\text{Validity} = \frac{\text{hits}}{\text{hits+false alarms}}$$

$$\text{Effectiveness} = \text{Thoroughness} \times \text{Validity}$$

Figure 1: Measuring Method Effectiveness

Discount methods certainly eased industry uptake of HCI approaches in the 1990s. However, the real determinant of appropriateness is not "discountability," per se, but rather cost-benefit, and must include not only an assessment of benefits but also of risks, especially risks of errors. In our research, we have used the concept of Effectiveness as developed by Sears (1997). Figure 1 shows the definitions and formulae used in Sears' measures, which address two kinds of error: missed usability issues and false alarms. Our research has led us to the conclusion that discount methods may be so error-prone that they discredit usability practitioners, and should be cleared off the HCI store's shelves.

Three ways to get usability on the cheap

To evaluate discount methods, we have to look at what is being traded-off in order to reduce costs, and consider the risks. Let us examine these tradeoffs in three broad categories of cost cutting tactics.

Cost Cut Tactic 1: Reduce the range of factors considered

Understanding usability problems requires attention to three key facets: (1) the contexts in which they arise, (2) the actual immediate and eventual difficulties, and (3) the assumed cause(s) of these difficulties (Lavery and Cockton 1997). Discount methods operate by reducing the facets under consideration. Inspection methods necessarily concentrate on causes. However, inspection methods cannot find actual difficulties, and do not pay attention to contexts. Discount-testing methods avoid the experimental controls that can confidently establish causation. Therefore, uncertainty over real difficulties or actual causes inevitably has an impact on the quality of recommended solutions.

Discount inspection methods save on time and skill by reducing the theory space for potential usability problems. Put simply, they narrow the scope of what the analyst has to consider. With less to look at and think about, analysts can work more quickly. Thus in Cognitive Walkthrough (Wharton et al. 1994), causes of potential problems are limited to "labels" that are hard to find or interpret, given a user's assumed knowledge. In Heuristic Evaluation (HE, Nielsen 1994), the theory space is limited to (classes of) system features that can cause problems. No discount method takes analysts systematically through a search space. Analysts must essentially pick sample user tasks or system features at random.

In discount user testing, limited user differences and data collection instruments restrict the range of difficulties that can be

recorded and/or reliably analyzed, as well as severely limiting consideration of the contexts under which difficulties can arise in the first place.

User difficulties result from a complex interaction of user and system factors. Strengths in one may compensate for weaknesses in others. So, expert or highly motivated users may be able to overcome design problems. Conversely, user incapability may not lead to usability problems if the system is not demanding. Discount methods are generally too simple to take such complex interactions into account. One result is false alarms, such as failing to realize that a system defect is neutralized within a particular interaction context (e.g., misleading status bar messages have no effect when users never read them!). Equally, a failure to consider complex interaction contexts can lead to problems being missed, for example when task breakdown occurs following several previous seemingly harmless user actions.

Inspection methods do not encourage analysts to take a rich or comprehensive view of interaction. Too often, most system features and user tasks get ignored, as does consideration of likely user knowledge or capabilities. Worst of all, inspection methods very rarely lead analysts to consider how system, user and task attributes will interact to either avoid or guarantee the emergence of a usability problem. Similarly, discount user testing inevitably restricts the range of user capabilities, knowledge and tasks sampled, and may similarly fail to expose test users to the system features that are most likely to result in unsatisfactory interaction.

One discounting tactic that has been advocated for user testing has been restricting the number of users tested to five or fewer. This will inevitably reduce user differences. Furthermore, many problems can often still be found with additional users (Spool and Schroeder 2001), even for a small subsystem. Figure 2 shows the results of a study involving 12 users (Woolrych and Cockton, 2001). Random selections of even six participants would result in widely differing views on the existence, frequency and severity of problems.

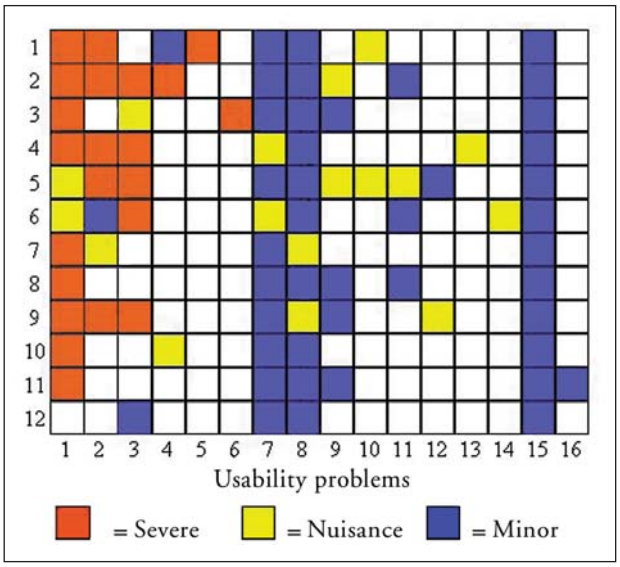

Figure 2: Problem Severity by Test Participant.

For example, contrast problems for Participants 1 to 6 with those for Participants 5, 6, 8 and (the more expert) 10 to 12. The testing order was due to participant availability, so the yield from the 'first' six is accidental.

Cost Cut Tactic 2: Pint sized methods inside a big box

Discount stores often offer small goods in large boxes. It looks like you are getting more than you actually are. Discount methods can do the same thing. In our research on HE, predictions attributable to HE rattled around within a big box of predictions based on analysts' common sense. We knew common sense was at play, since all analysts read training materials containing conformance questions for appropriate heuristic applications, and had to base predictions on these questions. We could thus code cited heuristics in reports as (in)appropriate. We found that, overall, 69 percent of predictions were associated with inappropriate heuristics. Furthermore, thoroughness of evaluations was mostly attributable to the individual skills of different analyst groups (Cockton and Woolrych 2001). Thus, it appeared that the analysts themselves, rather than the heuristics they were supposedly applying, provided the discovery resource. Interestingly, additional analysis indicated that hits associated with correct applications of the heuristics tended to be problems that had minor impact and/or low frequency. High

frequency/severity problems may simply be more "obvious" to analysts based on common sense, leading them to use arbitrary heuristics as post hoc justifications for the most critical usability problems. This can only impair the quality of recommendations for resolving a usability problem. Indeed, another weakness of HE is that it does little to support analysis of problem causes, leading to inappropriate solution generation.

from them? Often this is left to the consumer of the report. Without standard report formats, merging the predictions of individual analysts can be a frustrating experience (Connell and Hammond 1999). With a standard format, it remains time consuming and requires skill. Analysts' meetings to jointly prioritize problem predictions are effective, but this again requires time and skill, undoing any potential cost savings of the so-called discount

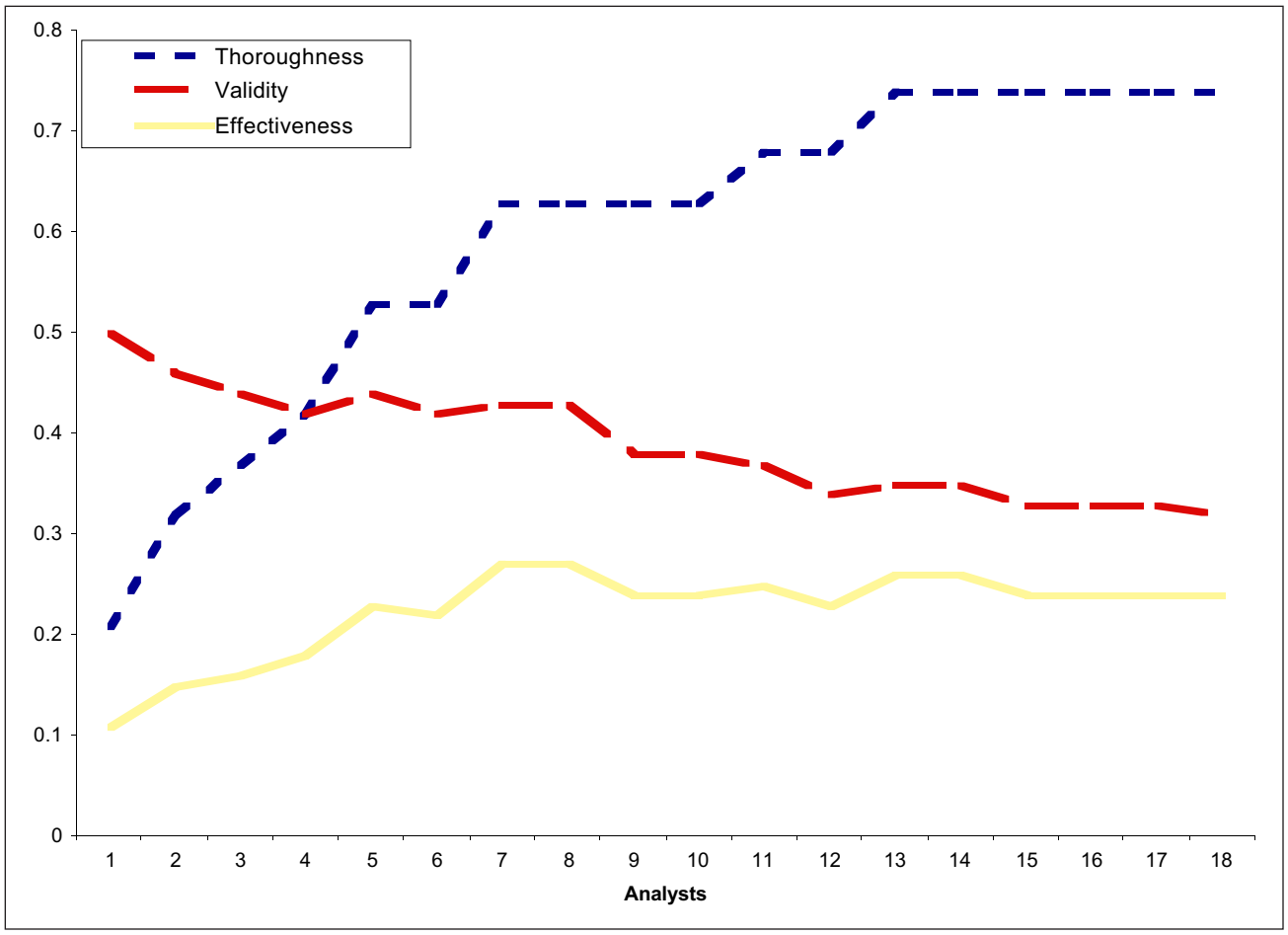

Figure 3: Cumulative Thoroughness, Validity and Effectiveness

Cost Cut Tactic 3: Self-Assembly Problem Sets

With self-assembly furniture, although the price may be reduced, the cost is not. The time that the purchaser spends assembling the furniture may be worth more than the price discount. The same is often true of discount usability methods, especially when multiple methods or HE carried out by multiple analysts are used. Who integrates and prioritizes the predictions and draws the conclusions

method. You never get anything for nothing.

Using multiple analysts may not be a safe way to compensate for the weakness of discount methods. Multiple analysts improve thoroughness because it only takes one analyst to discover a problem for it to be predicted. The impact on validity is less positive—it only takes one analyst to not eliminate a problem for it to not be eliminated. So, without a consensus-based prioritized master list, multiple

analysts will reduce a method's effectiveness (Woolrych and Cockton 2002).

Figure 3 shows data from our study (Cockton and Woolrych 2002) on how thoroughness, validity, and effectiveness change as analysts are added. Thoroughness was asymptotic (Landauer and Nielsen 1993), but validity got worse as analysts were added. The effectiveness trend was most interesting. It peaked at seven analysts, remained unaffected at eight, and then declined slightly, due to the negative impact of declining validity.

If analysts cannot be brought together to form a consensus, then perhaps simple frequency based elimination (as used in some method assessments) may help. Our study suggests that this is not necessarily true. There were thirteen unique predictions, including nine false alarms and four hits (Cockton and Woolrych 2001). When eliminated on the basis of prediction frequency, thoroughness and validity and effectiveness would all drop. Analyst consensus is safer than independent frequency-based elimination. However, analyst groups (or a more expert evaluation manager) could still eliminate actual problems or preserve false alarms.

The challenge is to improve all HCI methods, so that discount methods are less discounted and "full strength" methods can be applied in more contexts.

What does all this mean for our field?

Our recommendations for practice reflect input from colleagues at HCI 2001 (Ken Dye/Microsoft, David Roberts/IBM Warwick) and on CHI-WEB (AmandaPrail/Netusability, Fraser Hamilton/IconMediaLab, Josh Paluch/Ovo Studios).

First, there will probably always be a place for discount methods. The challenge is to improve all HCI methods, so that discount methods are less discounted and "full strength" methods can be applied in more contexts. Discount methods must become more effective and other methods must become more practical.

Second, discount methods are most appropriately used to drive design iterations, as opposed to providing summative evaluation, benchmarking or competitor analysis. However, even here they are risky. In most cases, a little more planning, better analysts, more users and more analysis will all pay off.

Third, participants are only one cost in user testing. Planning and analysis generally take more time than testing, and the difference in cost between five and 10 users can be relatively low. Clients on a limited budget have reduced costs by carrying out some planning themselves and by having developers attend during testing (thus reducing analysis costs). Look at the real costs of user testing and know where the costs originate. Try to find cost savings in planning and analysis as well as on participants. For inspections, too, look for ways to reduce hidden costs, such as problem merging.

Fourth, we must acknowledge that our studies only look at prediction effectiveness, and not at method impact. In real working contexts, impact comes not from usability experts generating solution recommendations in isolation, but from working together with multidisciplinary project teams to generate solutions. However, it seems fair to say that prediction effectiveness should be considered a prerequisite for impact effectiveness.

Fifth, the value of discount methods as training devices should not be underestimated! One valuable outcome of collaborative inspections may well be that the developer team will see that user testing is essential.

Sixth, errors arising from discount methods

BUSINESS COLUMN EDITORS
Susan Dray &
David A. Siegel
Dray & Associates, Inc.
2007 Kenwood Parkway
Minneapolis, MN 55405,
USA
+1-612-377-1980
fax: +1-612-377-0363
dray@acm.org
david.siegel@acm.org

may be more costly in some contexts than others. Different business models make different demands. In some contexts, hits may always be wins irrespective of the misses. However, in contexts such as online shopping, misses can be fatal. Savings on support costs and a few more attractive features are a benefit for retailed software, but for free-use software on the Web, it may be vital to eliminate *all* severe problems. Once software is bought, most users will (have to) struggle on with it. This is not true of free Web applications such as e-commerce sites. In general, discount methods are unable to address the whole product/site experience that is a key concern to DotCom managers.

Conclusions

Don't believe everything you read on the Web! Discount methods aren't very safe. They can and should be improved. Research has a key role here. In the meantime, understand method risks and do what you can to mitigate them.

PERMISSION TO MAKE DIGITAL OR HARD COPIES OF ALL OR PART OF THIS WORK FOR PERSONAL OR CLASSROOM USE IS GRANTED WITHOUT FEE PROVIDED THAT COPIES ARE NOT MADE OR DISTRIBUTED FOR PROFIT OR COMMERCIAL ADVANTAGE AND THAT COPIES BEAR THIS NOTICE AND THE FULL CITATION ON THE FIRST PAGE. TO COPY OTHERWISE, TO REPUBLISH, TO POST ON SERVERS OR TO REDISTRIBUTE TO LISTS, REQUIRES PRIOR SPECIFIC PERMISSION AND/OR A FEE.
© ACM 1072-5220/02/0900 $5.00

References

1. Cockton, G. and Woolrych, A. (2001). "Understanding Inspection Methods: Lessons from an Assessment of Heuristic Evaluation," in Blandford, A. and Vanderdonckt, J. and Gray, P. (eds.), *People and Computers XV,* Springer-Verlag, 171-192.

2. Cockton, G., Lavery, D. and Woolrych, A. (2002). "Inspection-based Evaluations" in *The Human-Computer Interaction Handbook*, eds. J, Jacko and A. Sears (in press) Lawrence Erlbaum, 1120-1140.

3. Connell, I. W. and Hammond, N. V. (1999). "Comparing Usability Evaluation Principles with Heuristics: Problem Instances vs. Problem Types," in Sasse, M. A. and Johnson, C. (Eds.), *Proc. INTERACT '99*, IOS Press,. 621-629.

4. Landauer, T. K. and Nielsen, J., (1993) "A Mathematical Model of the Finding of Usability Problems," in *Proceedings of INTERCHI '93,* ACM, 206-213.

5. Lavery, D. and Cockton, G. (1997),"Representing Predicted and Actual Usability Problems," in Johnson, H., Johnson, P., and O'Neill, E. (Eds.), *Proceedings of International Workshop on Representations in Interactive Software Development,* Queen Mary and Westfield College, University of London, 97-108.

6. Nielsen, J. (1994). "Enhancing the Explanatory Power of Usability Heuristics," In Adelson, B., Dumais, S., and Olson, J. (Eds.), in *Proc.CHI'94,* ACM, 152-158.

7. Sears, A., (1997) "Heuristic Walkthroughs: Finding the Problems Without the Noise," *International Journal of Human-Computer Interaction,* 9(3), 213-23.

8. Spool, J. and Schroeder, W. "Testing Websites: Five Users is Nowhere Near Enough. in *CHI 2001 Extended Abstracts,* ACM, 285-286

9. Wharton, C., Rieman, J., Lewis, C., and Polson, P. (1994). "The Cognitive Walkthrough: A Practitioner's Guide," in Nielsen, J. and Mack, R. L. (Eds.), *Usability Inspection Methods,* John Wiley and Sons, 105-140.

10. Woolrych, A. and Cockton, G., "Why and When Five Test Users aren't Enough," in *Proceedings of IHM-HCI 2001 Conference*, eds. J. Vanderdonckt, A. Blandford, and A. Derycke, Cépadèus Éditions: Toulouse, Volume 2,105-108, 2001

11. Woolrych, A. and Cockton, G., "Testing a Conjecture based on the DR-AR Model of Usability Inspection Method Effectiveness," to appear in *Proc. HCI 2002 Conference*, eds. H. Sharp et al., Volume 2, British Computer Society, London, 2002.