Changing Analysts' Tunes: The Surprising Impact of a New Instrument for Usability Inspection Method Assessment

Gilbert Cockton, Alan Woolrych, Lynne Hall, Mark Hindmarch
School of Computing and Technology, University of Sunderland, PO Box 299, Sunderland, SR6 0YN

Tel: **+44 191 515 3394**
Fax: **+44 191 515 2781**
*Email:****{Gilbert.Cockton,Alan.Woolrych,Lynne.Hall, Mark.Hindmarch}@sunderland.ac.uk***

We describe the impact on analyst performance of an extended problem report format. Previous studies have shown that Heuristic Evaluation can only find a high proportion of actual problems (thoroughness) if multiple analysts are used. However, adding analysts can result in a high proportion of false positives (low validity). We report surprising interim results from a large study that is exploring the DARe model for evaluation method effectiveness. The DARe model relates the effectiveness of an evaluation method to evaluators' command of discovery and analysis resources. Previous work has shown that Heuristic Evaluation poorly supports problem discovery and analysis: heuristics tend to be inappropriately applied to problem predictions. We developed an extended problem report format to let us study analyst decision making during usability inspection. Our focus was on the quality of insights into analyst behaviour delivered by this extended report format. However, our first use of this format revealed unexpected improvements in validity (false positive reduction) and appropriate heuristic application. We argue that the format has unexpectedly led to more care and caution in problem discovery and elimination, and in heuristic application. Evaluation performance can thus be improved by indirectly 'fixing the analyst' via generic fixes to inspection methods. In addition, we provide the first direct evidence of how evaluators use separate discovery and analysis resources during usability inspection.

Keywords: DARe (DR-AR) Model, Usability Inspection Methods, Heuristic Evaluation, Usability Evaluation, HCI Research Methods.

1 Introduction: The Discovery and Analysis Resource (DARe) Model for Usability Evaluation

The Discovery and Analysis Resource (DARe) model for Usability Evaluation distinguishes finding from keeping for usability problems. Analysts find *possible* problems (via inspection or user testing) and then either confirm them as *probable* problems (for user testing, significant), or eliminate them as *improbable* (insignificant). Analysts thus *discover* problems and then *analyse* them in separate phases of evaluation. Evidence for the DARe model and the range of resources used by usability analysts was derived from a large study (Cockton and Woolrych 2001, Woolrych 2001). The predictive power of the model was demonstrated in a subsequent re-analysis of the study data (Woolrych and Cockton 2002).

In this paper we report the first study based directly on the DARe model. Ten groups of final year HCI students applied the standard Heuristic Evaluation (Nielsen 1994) to a local transport web-site (http://www.tyneandwearmetro.co.uk/). They reported problems using an extended version of our problem report format (Cockton and Woolrych 2001), which required them to state their discovery tactics and their reasons for heuristic use and problem elimination or confirmation. The results provide firm evidence for distinct discovery and analysis resources in usability inspection, as well as clear-cut examples of specific resources in use. For the first time, we can isolate and analyse false negatives as well as false positives.

Despite the inherent limitations of self-reporting as a research instrument, we were surprised by unexpected impacts on analyst performance. Analyst application of Heuristic Evaluation was more valid (fewer false positives) and appropriate, and (as yet) no less thorough. We discuss how explicit reporting of discovery tactics and confirmation/elimination rationales can produce highly desirable improvements in analyst performance.

Before presenting the current study, we will argue that evaluator skills are a key variable in both analytical and empirical evaluation, and that neither class of evaluation has any automatic advantages or disadvantages. Potential method benefits can only thus be ensured by careful planning and skilled execution. We introduce the DARe model as a framework for structuring planning, skill development and reflective professional self-evaluation. We then present the main study and its results, showing how a simple extension to problem reports can improve the effectiveness of multiple analysts in Heuristic Evaluation.

2 Predictive Models in Usability Evaluation

Usability evaluation methods divide into two key groups: analytical and empirical. Empirical methods (Dumas 2003) observe systems/prototypes in use. Analytical methods examine systems (perhaps via models or specifications) to identify potential (ideally, probable) usability problems. The two main approaches here are *model-based* (Kieras 2003) and *inspection methods* (Cockton *et al*. 2003). The

inspection method assessed in this study is *Heuristic Evaluation* (Nielsen 1994). Both main groups of methods have potential advantages and disadvantages. However, these are not as clear-cut as often stated.

It is all too easy to make crude and shallow comparisons of analytical and empirical evaluation. The key common variable in both cases is the skill of usability specialists. Significant evaluator effects can be shown for both analytical (Cockton *et al* 2003) and empirical (Hertzum and Jacobsen 2001) methods. There are no absolute differences in cost or quality between the two types of method. Analytical methods are assumed to be faster, cheaper and more flexible. Empirical methods are assumed to be more reliable. None of these commonplace beliefs stand up to examination.

All evaluations have distinct phases of planning, implementation, analysis and recommendation. Reporting can occur in all phases (test scripts, session notes, problem reports, recommendations). Planning is not automatically faster or cheaper for analytical methods, although inspection methods require less effort than model-based ones (no models to construct). Both analytical and empirical methods require system familiarisation, and contextual understanding (users, goals, tasks, scenarios, domain knowledge). Empirical methods do require user recruitment and test protocol design, but these can be kept extremely simple (e.g., opportunistic 'hallway' testing, free usage, field observation). Inspection methods may require challenging recruitment and scheduling of multiple analysts.

Similarities in cost and speed apply to implementation. Managing a team of multiple analysts could be as time consuming as running several users through tests. Also, during analysis, merging analyst predictions can become as time consuming as analysing test data (especially for vague or conflicting predictions — Connell and Hammond 1999). Indeed, if developers are present during testing, analysis can be speeded and simplified, and even combined with agreement on necessary changes. Both method groups present challenges to speed and economy. Predictions may be too vague to support confident and detailed recommendations. User difficulties in tests may require extensive causal analysis in order to adequately ground recommendations.

Flexibility is also seen as an advantage for analytical methods, which do not require executable robust prototypes. However, neither do empirical methods, which have been applied successfully to a range of low fidelity prototypes in participative design approaches (Kuhn and Muller 1993).

Empirical evaluation can thus be fast, flexible and cheap (but at the expense of reliability — Woolrych and Cockton 2001, Barnum *et al.* 2003). Analytical evaluation can be slow, inflexible and expensive, and with no corresponding increase in overall effectiveness. Thus while increased resources *can* readily result in more reliable empirical evaluation (Woolrych and Cockton 2001), this is less likely with analytical evaluation. Inspection methods are thus almost inescapably *discount* methods, because existing investment strategies (e.g., multiple analysts) have inherent limits. Thus, extra analysts uncover more problems, but soon come to add *even more* false alarms (Woolrych and Cockton 2002). Each extra analyst also adds to the cost of problem set consolidation and pruning. Costs keep rising and returns soon drop.

This paper addresses the challenge of increasing the effectiveness of usability inspection methods (UIMs) without increasing resource costs per evaluation. The aim is to reduce the penalties of using inspection methods in a wide range of situations where user testing is infeasible or undesirable, for example, as an input to user test planning, for driving early design iterations, or for informing change decisions in response to user testing. Usability evaluation cannot be wholly empirical, and thus analytical methods must be made more effective.

3 A More DAReing Investment Strategy for UIMs

There are two possible responses to poor UIM quality other than the discount approach of stacking analysts high and selling methods cheap. Fewer and better skilled analysts could be used (and we should be able to determine analyst skill levels). Alternatively, better methods could be developed. The choice then is: fix the analyst or fix the method? Which is better?

The *Discovery and Analysis Resource* (DARe[1]) model suggests that fixing the analyst will be more effective than fixing the method. This follows from its modelling of the middle phases of usability evaluation: implementation and analysis. The DARe model identifies distinct knowledge resources that help analysts to find possible problems (discovery resources) and others that support either confirmation of probable (or elimination of improbable) problems (analysis resources). Our first study showed that Heuristic Evaluation (HE) was never clearly a discovery resource and typically not an analysis resource. Heuristics were not being used to find possible problems, as most (61%) heuristic applications were inappropriate. 80% of the hardest to construct problems were missed, again indicating that heuristics would not lead analysts to several serious problems. Nielsen's 1994 set of ten heuristics is derived from seven factors that could only predict 30% of an eleven system corpus of 247 usability problems, so it is hard to believe that our analysts' prediction of 72% of actual problems was due to the disclosing power of heuristics. As Gray and Salzman (1998) have commented, Nielsen's studies allow nothing to be attributed to HE, instead for example, in one study (Nielsen 1992), it was experts alone who found more problems than novices, and *not experts using HE*.

Similarly, analysts in our first study failed to eliminate far too many false positives. 65% of their predictions did not transpire in a carefully designed falsification test. HE thus failed to eliminate a host of improbable problems (and many bogus ones too). Overall, HE clearly played a limited role in the discovery of possible problems and the elimination/confirmation of im/probable ones. This let us derive a conjecture that multiple analysts can improve discovery resources (since these are additive), but that they will dilute elimination analysis resources (since one bad apple can spoil the bunch), and thus increased thoroughness (hit rate) will be at the expense of lowered validity (false positive rate), and thus overall effectiveness (thoroughness x validity) will be reduced if losses outweigh gains.

[1] In previous publications, this has been called the DR-AR model (Discovery Resources — Analysis Resources). The reason for the change should be clear.

This turned out to be true for a retrospective analysis of our first study's data (Woolrych and Cockton 2002).

Given the importance of analyst resources in method effectiveness, it appeared that investing in analysts would have a much more immediate and reliable payback than investing in methods. We could simply see no way on improving on Nielsen's attempted grounding of heuristics in a usability problem corpus (Nielsen 1994). HE, as far as we could see, was beyond repair. Analysts however, like all humans, could still be saved.

The DARe model suggests that making analysts aware of effective discovery and analysis resources could encourage them to apply knowledge resources consciously during inspections. However, to test this conjecture on the value of reflective use of known resources, we needed a better understanding of their nature. The first study impeded this in two key ways. Firstly, we derived the DARe model from that study, and thus were unprepared for separating the use of a knowledge resource for discovery from one for analysis. For example, knowledge of user tasks could lead analysts to discover a problem, but it could equally have been used to confirm it. Secondly, we had no evidence whatsoever of elimination analysis. We could see how and why many false positives *should* have been eliminated, but we had no way of seeing how any problem was confirmed or eliminated. In particular, we could not study false negatives, since HE, unlike Cognitive Walkthrough (Wharton *et al.* 1994), has no success cases that would expose erroneous exclusion of probable problems.

We had thus derived the DARe model from failures of elimination and discovery (and failure of HE to be involved in discoveries). The existence of confirmation was a logical dual to elimination, but again, we had no way of actually distinguishing discovery resources from confirming analysis resources. We thus designed the current study as an initial direct investigation into the nature of discovery and analysis resources in UIMs. To see resources in action, we had to extend the research instruments beyond those used in our original study.

4 A New Instrument for the Assessment Ensemble

UIM assessment requires multi-instrument research protocols. A key instrument is the structured report format (Lavery *et al.*, 1997, Lavery and Cockton 1997), which eases subsequent merging into a single predicted problem set. Other instruments such as analyst debriefings (individual, groups, plenary) can extend or refine this problem set. Actual problem set elicitation requires further research instruments such as video recording and debriefing interviews. Ideally, test data should be systematically analysed to produce problem reports identical in format to those used by analysts. The SUPEX method (Cockton and Lavery 1999) supports such analysis.

In the current study, we extended the report format to let analysts self-report on discovery resources and confirmation/elimination rationales. A key methodological aim of the study was to establish the limits of self-reporting in this setting (theoretically, there must be limitations; practically, these must be clearly identified). However, the primary aim of the study was to gather clear evidence on

the use of discovery and confirmation/elimination resources. We did not expect the new report format to radically alter analyst performance. However, the measures used in the initial study (thoroughness, validity, appropriateness) would be reapplied routinely to the current study, readily revealing performance changes.

4.1 Method

The current study follows the approach in the original study of comprehensive heuristic evaluations, using carefully designed falsification tests to validate analyst predictions (Cockton and Woolrych 2001, Figure 2). The user tests not only address the confirmed predictions, but also ones discovered but subsequently eliminated by analysts. We thus had to develop a report format that recorded both confirmed predictions and also discoveries 'discarded' following analysis.

The report format required a more detailed record of usability problem discovery and analysis. The problem report format had four main report sections. Part 1 requires analysts to describe the problem and associated user difficulties, using the same format as the initial study. Its purpose is to positively identify the usability problem reported, with four elements providing multiple points of reference for problem merging:

PROBLEM DESCRIPTION
The analyst must provide a brief description (in their own words) of the problem.

LIKELY/ACTUAL DIFFICULTIES
The analyst must record the anticipated difficulties the user will encounter as a consequence of the problem.

SPECIFIC CONTEXTS
The analyst is required to identify any specific contexts in which the problem may occur.

ASSUMED CAUSES
The analyst should describe the cause(s) of the problem, in their own judgment.

Part 2 addresses discovery resources and methods. Two general issues are covered. Firstly, analysts must record any reflection on individual problem discovery. The purpose is to identify what method resources (if any) assisted problem discovery. HE prescribes no particular strategy for system inspection (so in what sense is it an inspection *method?*). Secondly, analysts must classify their adopted discovery method as one of four categories. These are ordered by analyst effort in terms of planning and control over the inspection processes. The ordering starts with the easiest and ends with the most onerous method:

1. SYSTEM SCANNING – analysts simply 'looks around the system' with no particular strategy.

2. SYSTEM SEARCHING – a basic strategy (e.g., inspecting various links, toolbars or menu options) — effectively structured scanning.
3. GOAL PLAYING – involves role playing a specific user goal, for example, finding a specific piece of information.
4. METHOD FOLLOWING – goal playing, but walking through a preconceived method.

Part 2 also asks analysts to provide a confirmation rationale for probable problems. Part 3 of the report deals specifically with heuristic application to individual problems. Analysts must provide evidence of conformance of heuristics rather than just cite a heuristic relevant to individual performance.

Part 4 requires analysts to justify any problem elimination. The analyst is requested to clearly state why any problem initially discovered should warrant elimination, with specific reference to user impact.

31 undergraduate analysts from a final year HCI course worked in ten groups (one pair, seven groups of three and two of four) to complete a HE of a local transport web-site (http://www.tyneandwearmetro.co.uk/). All analysts used the extended problem report format. Predictions were merged into a single predicted problem set. We could then apply appropriateness measures to the set, as well as investigate the relationship between discovery methods and appropriateness, and discovery methods and elimination rates.

Appropriateness analysis followed the approach described in (Cockton and Woolrych 2001). Briefly, appropriate heuristic applications can be determined by correspondence between predicted difficulties and/or assumed causes and applicability criteria as stated in a HE training manual (Lavery *et al.*, 1996).

To compute other key measures such as thoroughness and validity, an actual problem set (typically from user testing) is required. Thoroughness and validity are defined:

$$\text{Thoroughness} = \text{hits} / (\text{hits} + \text{misses})$$

$$\text{Validity} = \text{hits} / (\text{hits} + \text{false positives})$$

where a *hit* is an actual problem matched by predictions, a *miss* is one unmatched by any predicted problem, and a *false positive* is a predicted problem that is not part of the actual problem set derived from user testing.

To plan user testing, two of the authors applied a card sort to the predicted problem set with the aim of isolating common site features and task steps. Falsification test scripts were then derived to systematically expose test users to site features and task steps that were predicted to be the causes or contexts of likely user difficulties. These test scripts were piloted (two obscure feature groups were initially excluded) with two test users. Test scripts were then revised to focus on unconfirmed and initially excluded problems. We thus used a simpler script to establish a core of successful predictions and then refined and extended the script to focus attention on predictions that had not been immediately confirmed. Three more test participants used the re-focused test script. The users were aged between 21 and 34, two males and three females. All had good computer/web literacy,

although the youngest claimed limited web experience. As a result of these five users' tests, 20 actual problems were found.

A constant focus on unconfirmed predictions is essential to ensure correct ultimate coding of some predictions as false positives. This is the basis of *falsification testing* — the main aim is to maximise confidence in false positive coding. Confidence in thoroughness scores (correct predictions/all actual problems) is secondary and these are thus always maxima, i.e., further testing would reduce thoroughness scores by increasing the actual problem set. However, it should be impossible to identify how further testing could convert a false positive to a correct prediction by finally exposing the predicted problem. If this is possible, then the test script must be revised and used by further test participants.

Test user problems were either matched to predictions (hits) or to discovered problems eliminated in error (false negatives), or were added as unpredicted problems (misses) at the end of each test session. One researcher conducted the user testing, and another researcher coded problems as they arose. Post test analysis revisited problem matches and additions. Thus we could recalculate thoroughness and validity after each user test, as predictions were matched and missed problems emerged. The SUPEX method (Cockton and Lavery 1999) was not applied to test problem extraction in this or the previous study. For this study, matching was performed by two of the authors, with further checks by the other authors. In theory, improved validity could be due to more generous matching of predicted to actual problems in this study over the previous one. We need to later exclude this possibility to draw any sound conclusions from this study.

4.2 Issues with Study Comparisons

Out of curiosity, we compared appropriateness, validity and thoroughness scores at this point in the current study with those from our initial large assessment of Heuristic Evaluation. We were surprised by what we found. In order to ensure comparability, we reanalyzed the initial study's data, which contained predictions from a coached single analyst and a pair of visiting masters degree students who used a combination of inspection methods. We reduced our predicted problem set to contain only predictions by undergraduate groups. This reduced the 99 analysts in 18 groups (Cockton and Woolrych 2001) to 16 groups of 96 analysts. We recalculated the scores for thoroughness (drops from 0.74 to 0.63 with the removal of only three coached analysts), validity (drops from 0.35 to 0.31) and appropriateness (drops from 39% to 31%). The 31% is for all predictions, and not just for hits only as the 39% in (Cockton and Woolrych 2001).

Thus we began by recalculating comparable scores. However, the two studies were carried out three years apart with different groups of final year undergraduates (10 groups of 31 as opposed to the initial 16 groups of 96), on different applications (drawing tool vs. web-site) and with mostly different test users (fifteen versus five). For the comparisons of thoroughness, validity and appropriateness, the possibility of significant confounds must be excluded.

Thoroughness scores could be biased if usability problems were easier to predict for an application, misses are missed (e.g., due to mostly expert test users)

or analysts are smarter. We have no evidence of the latter. Concerns over missed misses are reduced by a similar actual problem set (20, initial study = 19), so differences in overall test participant expertise have not resulted in fewer actual problems. As for ease of prediction, the fifteen hits for this study exceed the twelve for the initial study (coached analysts excluded), so if there is evidence in problems being easier to find for one application, then the current study benefits. However, we make no claims below for improved or reduced thoroughness. Still, the hit rate does impact validity scores.

For *validity*, analyst and test user differences between the two studies could result in bias. Again, we have no evidence that the two analyst cohorts differ in skills, and further test participants can only increase hits and thus validity (by converting false positives to hits). As long as we claim an *increase* in validity, further user testing could not undermine this. The issue is whether this is due to improved thoroughness, reduced false positives, or some combination of the two.

Validity scores depend on the hit rate and the false positive rate. Hits can only increase. However, with 20 problems revealed by five users (initial study: 19 problems by 15 users), we feel that significant increases are unlikely. Still higher thoroughness will partly explain improved validity. We must thus separate the impact of hits and false positives. As noted, the latter can only improve (i.e., decrease), so any improvements in validity reported below will be a minimum.

We also need to exclude the argument that fewer groups and analysts would inevitably result in fewer total false positives. This may be true for the complete predicted problem set, but if there is a significant drop in false positives per analyst group, then this argument will not stand.

For *appropriateness*, only analyst cohort differences could bias results. They received the same training on HE: a very similar lecture and access to the same training handbook (Lavery *et al.*, 1996). Possible (but unlikely) student differences apart, the only difference between the studies here was the problem format, to which we thus could attribute any differences in appropriateness. Also, we again need to exclude the argument that fewer groups and analysts would result in fewer misappropriate applications overall. If there is a significant rise in appropriate heuristic applications per analyst group, then this argument will not stand. Nor will the argument stand that application differences between the two studies eased or hindered heuristic application, since Instone (1997) could easily illustrate the use of HE for web applications.

To conclude, while potentially confounding effects of inter-study differences are logically possible, some are no more than logical possibilities that cannot be shown to have transpired (e.g., hidden variable distorting discoverability distributions), while others could only improve the results reported in the next section (e.g., rise in validity due to further confirmation of predictions).

5 Results

Analysis of the completed problem reports resulted in an initial problem set of 37 discovered problems, of which analyst groups eliminated nine. Despite having under one-third of the analysts (31 as opposed to 96) and just under two-thirds (10

as opposed to 16) of the groups relative to the initial study, the total problem discovery is almost identical (37 as opposed to 40). Table 1 shows further comparisons between prediction set sizes per group in each study. There is no significant difference between these set sizes, adding confidence to the comparability of the two studies. Given that discovered problems could be eliminated in the current study, this shows that groups did not significantly reduce their predictions when using the extended problem format.

	Initial Study	**Current Study**
Smallest set	1	2
Largest set	15	5
Median size	5.5	4
Mean size	5.4	3.8

Table 1: Groups' Prediction Set Sizes

For largest prediction sets, two groups in the initial study 'outperformed' those in the current study. They each predicted a unique problem, and thus it could be argued that thoroughness could have been higher had we used more groups. Equally though, it could have been lower if we had tested more users.

Six problems were identified during user testing in addition to the fourteen confirmed by analysts. This gives a thoroughness of 0.7, compared to 0.63 for the student groups in our initial study. Given that we have reported when and why five users aren't enough (Woolrych and Cockton 2001), we will make no claims on thoroughness until we have tested more users, despite a current difference of over 10% between the two studies. While it appears that improved validity is not at the expense of reduced thoroughness, and thus overall effectiveness (thoroughness x validity) is improved, testing a further 10 users to achieve comparability with our first study would inevitably lower thoroughness.

Table 2 compares thoroughness between the two studies. There are no significant differences here, indicating that any increase in validity will not be largely due to a raised hit rate.

	Initial Study	**Current Study**
Lowest score	0.05	0.05
Highest score	0.37	0.2
Median score	0.11	0.1
Mean score	0.14	0.12

Table 2: Groups' Thoroughness Scores

The highest thoroughness score in the initial study is worth examining. The next best score was 0.21, so much of the similarity between the two studies is due to a single group of tenacious analysts. However, their validity of 0.47 would have been the third lowest in the current study, so their success here came at a price.

There were thus no significant differences for either prediction rate (despite directing analysts to eliminate improbable problems) or thoroughness (despite directing analysts to more onerous discovery resources). Both of these add credibility to two surprising significant differences.

5.1 First surprising impact: false positives

Of the 28 confirmed predictions, 14 were matched to actual problems. This gives a validity of 0.5, compared to 0.31 for the student groups in our initial study. The proportion of false positives in the merged problem set, though still high, has dropped by almost 20%. More significantly, the mean count of false positives per group halved from 3.13 (initial study) to 1.5 (current study). The difference in false positive counts between the two studies was fairly significant (t-test, $p = 0.0176$). We can thus exclude the argument that fewer groups and analysts would inevitably result in false positives overall, as there is a significant drop in false positives per analyst group.

Few software developers would pass over such an improvement. The reason appears to be the elimination of improbable and bogus problems, which was generally successful (totally successful in the case of bogus problems, see below). To be eliminated from the merged predicted problem set, a problem had to be eliminated by all discovering groups. There were nine such eliminated problems, of which eight cannot be associated with actual user problems (correct elimination). However one did arise in user testing, and was thus incorrectly eliminated. The ability to detect such false negatives is a valuable property of the extended report format.

A significant reduction in false positives per group did not result in a significant increase in validity per group. Table 3 compares validity scores for the two studies. They are similar until we exclude the best and worst 3 groups in the initial study (leaving the middle 10 groups). The mean drops considerably, since the loss of the worst performers cannot offset the loss of the best analysts. The failure of significantly lower false positives to be reflected in significantly higher validity is largely due to the performance in the initial study of two groups with a validity of 1 (1/1 and 4/4 predictions valid) and two with 0.67 (both 2/3 valid).

	Initial Study	**Current Study**
Lowest score	0.13	0.25
Median score	0.5	0.59
Mean score	0.52	0.59
Mean score (middle 10)	0.32	0.59
Highest score	1	1

Table 3: Group Validity Scores

Reduced false positives were an unexpected surprise, but further user testing is required to establish whether this could be at the expense of reduced thoroughness, and ultimately, reduced effectiveness. However, there already

appears to be a relationship between discovery resource usage and false positive elimination.

Analysts were far more likely to choose *system scanning* or *system searching* as their preferred method of identifying predictions (significant chi square at .01 level). A smaller number of predictions were made using *goal playing*. So few predictions were made using *method following* that the data relating to use of this method has limited use.

Analysts who used system scanning were more likely to keep (i.e., not eliminate) predictions, and confirmed 81% of their discovered possible problems. Analysts using *system searching* were most likely to eliminate predictions, eliminating 41% of their discovered possible problems. *System scanning* resulted in the least valid performance, with only 47% of the predictions currently confirmed by user testing. In contrast, 92% of discoveries identified through *goal playing* and confirmed as probable problems have been confirmed by user testing. So, validity for problems found via *goal playing* was double that of s*ystem scanning*, and further user testing is unlikely to significantly close this gap.

System scanning and searching were both associated with high numbers of false positives, whilst goal playing resulted in very few, suggesting that this discovery method can support significantly more valid performance. However, goal playing is not a method that is explicitly advocated for heuristic evaluation, whilst system scanning and searching are. General education on discovery method resources here could thus have improved analyst performance by making false positives harder to find! This may seem perverse, but our data does support this, and in a way that helps us to exclude matching bias (over generous matching of actual to predicted problems) as a confounding cause of the improved validity scores.

More structured and user-centred discovery methods were associated with a higher elimination rate, providing clear evidence for the removal of potential false positives (in fact, there was only one false negative as a result). However, even the system scanners showed one improvement over the first study's analysts. There were no bogus problems. Neither factually or logically bogus (Cockton and Woolrych 2001) predictions were made, i.e., there were no errors of fact or recommendations based on flawed design rationales in the problem set for this study. Given that 32% of the false positives in (Cockton and Woolrych 2001) were bogus, this may, combined with the clear evidence of correct eliminations, represent all of the reduction in false positives as a result of the extended report format. Our explanation is simply that the extended format makes it very difficult to report bogus problems, since reports of such problems would look very sparse in the extended format. The resulting seas of white space in a problem report probably discourages analysts from making simple false assertions that features don't exist (indeed, they instead reported correctly that they couldn't find them). The format also considerably obstructs the reporting of logically bogus recommendations, since these take the form *"I can think of a better option because <flawed rationale>"*, since the only way to get such a back to front problem report into the extended format is to leave all but the first part blank,

We are thus confident that, despite a low level of control over experimenter bias in extracting actual problems and matching them to predicted ones, the resulting data lets us exclude this as a systematic source of bias.

5.2 Second surprising impact: appropriateness

Appropriateness of heuristic application rose to 57% (from 31% for all student predictions in first study), a 26% practical improvement that would be welcome in all practical settings, since inappropriate heuristic applications could result in inappropriate recommendations for design changes. Only one group in the current study scored below the overall average for the initial study (where they would have been ranked seven out of sixteen). The difference in appropriateness between the two studies was very significant (t-test, $p = 0.0018$). We can thus exclude the argument that fewer groups and analysts would inevitably result in fewer misappropriate applications overall, as there is a significant rise in appropriate heuristic applications per group. Thus the overall drop in misappropriate heuristic use is due to improved performance across all groups, and not just a result of having fewer groups. Table 4 compares appropriateness for the two studies.

	Initial Study	**Current Study**
Lowest score	0%	20%
Median score	27%	65%
Mean score	31%	61%
Highest score	80%	100%

Table 4: Groups' appropriateness scores between studies

We were also able to examine the relationship between appropriateness scores and discovery methods. Only 39% of problems found by system scanning were associated with appropriate heuristics, a rate similar to that for our initial study. For system searching, this rose to 70%, and to 73% for goal playing and method following combined. This means that we cannot completely attribute improvements here to Part 3 of the extended format, which requires heuristics to be not just named, but justified alongside confirmation rationales. Appropriate heuristic use is associated with more onerous discovery methods, which would appear to contribute to the difference between the two studies.

5.3 Confirmation of the DARe model

We now have unequivocal evidence for the fit of the DARe model to analyst behaviour. Before the current study, we had only a post-hoc derivation of the model (Cockton and Woolrych 2001), with additional support from reanalysis of multiple analyst performance (Woolrych and Cockton 2002). We now have evidence that clearly identifies different discovery and analysis resource usage.

A striking example of different discovery and analysis resources shows how carefree discovery can be combined with more sober analysis:

> "After seeing the different looking button, I decided to click on it and explore further" (Group 6, problem 6)

The report confirmed that *system scanning* was used to find this possible problem, but the confirmation rationale made clear that the button was "unclear and misleading", with likely difficulties being confusion and inability to recover from taking this off-site link. The prediction was confirmed by user testing.

The key point is that system-centred *discovery* resources can be combined with user-centred *analysis* resources to produce successful predictions. Here, the analyst encountered a possible problem by "playing" with the system and then empathising with users to confirm the problem, correctly associating the problem with the heuristic *user control and freedom* (can't undo). Another associated heuristic, *consistency and standards*, was inappropriate. There were no inconsistencies here and it is not clear that there *is* an agreed one-size-fits-all web standard on off-site links (just good advice that turns out to apply here).

Further striking examples arise with eliminated problems. System scanning discovered the one false negative, which was eliminated (after consideration of accessibility issues) by assumptions about user capabilities (i.e., all users were as capable as the analysts!) User testing invalidated these assumptions. In contrast, a successful elimination of a possible problem (missing back-links from a questionnaire page) was eliminated by realising that all links into the page were from one level below the home page, and that users could thus be expected to (and did) find their way back. This possible problem was discovered by system scanning, and eliminated by knowledge of web interaction.

The availability of clearly separated discovery, confirmation and elimination resources lets us develop educational materials that we hope can improve analyst performance across a range of usability inspection methods.

6. Discussion

In both studies, groups made similar numbers of predictions. The current study's analysts are simply more valid (significantly fewer false positives per group) and use heuristics more appropriately. Improved validity can be reasonably attributed to the impact of the extended report format that encouraged analysts to consider more structured discovery methods, to eliminate improbable problems, and to avoid bogus predictions. Reduction in false positives can be attributed to discovering fewer improbable problems or to their elimination, which was, with one exception, well considered. Although validity is much improved, it remains poor. Continued over-reliance on system scanning appears to lead analysts to more improbable problems that are less likely to be eliminated. Easily found problems, it appears, are harder to lose!

The much improved appropriateness scores are due jointly to report format requirements for confirmation and justification, and for conscious discovery and elimination. Having taken most analysts through the pain barrier of reflection, self-criticism and rational argument, they become better prepared for thinking

carefully about appropriate heuristic usage. Indeed, as confirmation rationales are formed, analysts may find it easier to identify the most appropriate heuristic, where it exists.

It thus appears that we can fix the analyst indirectly via fixing the method. In this case, there was no change to the actual heuristics in use. Instead, the process of discovering and analysing problems was made more explicit. This approach could be applied to all inspection methods. In the case of Cognitive Walkthrough (Wharton *et al.* 1994), minimal extensions would be required to achieve this.

The extended problem report format is simpler than Sears (1997) approach to reducing false positives with HE, by prefixing a Cognitive Walkthrough. We would argue that we can achieve comparable results without the complexity of a two phase hybrid method, which we feel could adversely impact on thoroughness. Encouraging analysts to use more structured discovery methods and to explicitly confirm or eliminate problems may be enough.

7 Further Work

This paper is an initial report from a large iterative study that will continue user testing and analyst inspections in order to fully explore the DARe model for both empirical and analytical methods. We know that as additional analysts are added and more users are tested, that there will be changes to the scores reported above, especially thoroughness, for which we make no claims other than an initial lack of apparent reduction. However, the main results reported above will not change. We thus we have not rushed through 10 further user tests in order to achieve better comparability with the initial study. In terms of the DARe model, additional users are only one discovery resource. Further usability problems can just as easily be found by changing the test scripts, by letting users prepare their own tasks, by letting users just explore the system, by field studies or by web log analysis (Barnum *et al.* 2003). Changes to test protocols, including more extensive user debriefings and more active or less passive experimenter intervention can also increase the problem yield (albeit with concerns about reliability). We intend further user testing to be asymptotic, i.e., we will keep adding users and changing test protocols until we stop finding new problems.

We will thus test more than five users, but we will also make systematic changes to the testing procedures that cannot be 'rushed through'. We have therefore reported a surprising interim result with immediate implications for usability specialists: changing report format can improve analyst performance on validity and appropriateness. Establishing the impact on thoroughness requires further work as stated. Only tentative conclusions can currently be drawn here.

We will also further develop the report format to improve analyst competence and resource elicitation, and will explore analyst interviews and group discussions as for eliciting analyst use of discovery and analysis resources.

8 Conclusions

Improved evaluation performance occurs when analysts are required to explicitly report and rationalise their use of heuristics, and of discovery method and confirmation/elimination knowledge resources. A report format that demands more reflection appears to enhance usability inspection, resulting in fewer false positives and more appropriate heuristic usage. Neither of these claims appears to be undermined by differences between the current and earlier study. This suggests that approaches derived from the DARe model can significantly improve HCI methods that have stagnated for a decade.

UIMs can clearly be improved. However, even with the current report format, analysts continue to fail to find all problems and still generate false positives. There is room for improvement in discovery methods and analysis resources. Through focusing on method extensions, such as analyst education to improve competence, it becomes possible to fix the analyst with a fixed method.

The evolution of the DARe model shows the value of constant improvement in research protocols and instruments, with each iteration initially exposing and then extending its validity and applicability. At the same time, generic and specific improvements to inspection methods are developed. This theory driven approach shows the effective and valuable coupling of research and practice in HCI where pragmatic practice based approaches have failed to deliver method improvements.

References

Barnum, C., Bevan, N., Cockton, G., Nielsen, G., Spool, J., and Wixon, D., "The "Magic Number 5": Is It Enough for Web Testing?" in CHI 2003 Extended Abstracts, eds. G. Cockton et al., pp. 698-699, 2003.

Cockton, G. & Woolrych, A., "Understanding Inspection Methods: Lessons from an Assessment of Heuristic Evaluation," in People & Computers XV, eds. A. Blandford & J. Vanderdonckt, Springer-Verlag, 2001

Cockton, G., and Woolrych, A., "Sale must end: Should Discount Methods be Cleared off HCI's Shelves?", Interactions, XI(5), 25–30, ACM, 2002.

Cockton, G. and Lavery, D. "A Framework for Usability Problem Extraction", in INTERACT 99 Proceedings, eds. A. Sasse and C. Johnson, 347-355, 1999.

Cockton, G., Lavery, D., and Woolrych, A., "Chapter 57: Inspection-Based methods", The Human-Computer Interaction Handbook, eds. J. Jacko and A. Sears, 1118–1138, Lawrence Erlbaum Associates, USA, 2003.

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

Dumas, J.S., "Chapter 56: User-Based Evaluations", The Human-Computer Interaction Handbook, eds. J. Jacko and A. Sears, 1093–1117, Lawrence Erlbaum Associates, USA, 2003.

Gray, W.D. & Salzman, M., "Damaged Merchandise? A Review of Experiments that Compare Usability Evaluation Methods", Human-Computer Interaction, 13(3), 203-261 1998.

Hertzum, M. & Jacobsen, N.E., "The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods", International Journal of Human-Computer Interaction,13(4), 421-443, 2001.

Instone , K., "Usability Engineering on the Web," in Advancing HTML: Style and Substance, 2(1), available at http://www.w3j.com/5/s3.instone.html, accessed 10/2/03, 1997

Kieras, D., "Chapter 57: Model-based evaluations", The Human-Computer Interaction Handbook, eds. J. Jacko and A. Sears, 1139–1168, Lawrence Erlbaum Associates, USA, 2003.

Kuhn, S., Muller, M.J., "Participatory Design - Introduction to the Special Section," in CACM, 36(6), 24-28, 1993

Lavery, D. and Cockton, G., "Representing Predicted and Actual Usability Problems", in Proc. Int. Workshop on Representations in Interactive Software Development, QMW London, 97-108, 1997.

Lavery, D., Cockton, G., and Atkinson, M. P., Heuristic Evaluation: Usability Evaluation Materials, Technical Report TR-1996-15, University of Glasgow, 1996. Available at http://crete.dcs.gla.ac.uk/publications/reports/1996-15.pdf

Lavery, D. Cockton, G. and Atkinson, M.P.,"Comparison of Evaluation Methods Using Structured Usability Problem Reports," in Behaviour and Information Technology, 16(4), pp. 246-266. 1997

Nielsen, J. 1992. Finding usability problems through heuristic evaluation. *Proceedings ACM CHI'92 Conference* (Monterey, CA, May 3-7), 373-380.

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

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

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

Woolrych, A. Assessing the Scope and Accuracy of the Usability Inspection Method Heuristic Evaluation, MPhil Thesis, Univ. Sunderland, UK, 2001.

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

Woolrych, A. and Cockton, G., "Testing a Conjecture based on the DR-AR Model of UIM Effectiveness ," in Proceedings of HCI 2002, Volume 2, eds. H. Sharp, P. Chalk, J. LePeuple and J. Rosbottom., British Computer Society, 30-33, 2002.